Bayesian Multi-Population Haplotype Inference via a Hierarchical Dirichlet Process Mixture

Eric P. Xing† EPXING@CS.CMU.EDU
Kyung-Ah Sohn† KSOHN@CS.CMU.EDU
Michael I. Jordan‡ JORDAN@CS.BERKELEY.EDU
Yee-Whye Teh♯ YEEWHYE@GMAIL.COM
† School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213; ‡ Computer Science Division and Department of Statistics, UC Berkeley, CA 94720; and ♯ Department of Computer Science, National University of Singapore, Singapore 117543.

Keywords: Dirichlet process, clustering, Bayesian inference, haplotype inference, statistical genetics

Abstract

Uncovering the haplotypes of single nucleotide polymorphisms and their population demography is essential for many biological and medical applications. Methods for haplotype inference developed thus far—including methods based on coalescence, finite and infinite mixtures, and maximal parsimony—ignore the underlying population structure in the genotype data. As noted by Pritchard (2001), different populations can share certain portion of their genetic ancestors, as well as have their own genetic components through migration and diversification. In this paper, we address the problem of *multi-population haplotype inference*. We capture cross-population structure using a nonparametric Bayesian prior known as the hierarchical Dirichlet process (HDP) (Teh et al., 2006), conjoining this prior with a recently developed Bayesian methodology for haplotype phasing known as DP-Haplotyper (Xing et al., 2004). We also develop an efficient sampling algorithm for the HDP based on a two-level nested Pólya urn scheme. We show that our model outperforms extant algorithms on both simulated and real biological data.

1. Introduction

Recent experimental advances have led to an explosion of data which document genetic variation at the DNA level within and between populations. For example, the international SNP map working group (2001) has reported the identification and mapping of 1.4 million single nucleotide polymorphisms (SNPs) in the human genome. These kinds of data lead to challenging inference and learning problems, problems whose solutions could lead to greater understanding of the genetic basis of disease propensities and other complex traits (Chakravarti, 2001; Clark, 2003).

We concern ourselves with the following problems. For autosomal loci in the genome of diploid organisms, the SNP *haplotypes*, which correspond to the joint identities of a contiguous sequence of multiple SNPs, are inherently ambiguous when only their genotypes (the unordered set of SNP alleles) are given (Clark, 1990; Hodge et al., 1999). The problem of inferring SNP haplotypes from genotypes is essential for the understanding of genetic variation in a population. A straightforward statistical genetics argument shows that the problem of haplotype inference can be formulated as a mixture model, where the set of mixture components corresponds to the pool of ancestor haplotypes, or *founders*, of the population (Excoffier & Slatkin, 1995; Niu et al., 2002; Kimmel & Shamir, 2004; Xing et al., 2004). Crucially, however, the size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. Thus we have a mixture model problem in which a key aspect of the inferential problem involves inference over the number of mixture components.

This uncertainty regarding the size of the haplotype pool is an instance of the perennial problem of "how many clusters?" in the clustering literature. The problem is particularly salient in large data sets where the

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006.

number of clusters needs to be relatively large and open-ended—exactly the scenario in population genomic analysis. Model selection techniques based on fixing the number of clusters and using an information-theoretic score to gauge the appropriate number seem misplaced when the uncertainty is so high. An alternative is provided by nonparametric Bayesian models, specifically the *Dirichlet process* (DP) (Ferguson, 1973; Blackwell & MacQueen, 1973), which provides a prior and posterior distribution for mixture models with unbounded numbers of mixture components. Xing et al. (2004) proposed to address the haplotype inference problem by exploiting DP mixtures. Completing the DP model specification with a set of ancestor-specific inheritance models, this approach yields Bayesian posterior inference for many of the entities of interest in the haplotype problem, including ancestral states and haplotype phases. The performance of the DP-based haplotyper was shown to be comparable to the state-of-the-art haplotype inference algorithm, PHASE (Stephens et al., 2001), and it significantly outperforms other algorithms based on finite mixture models (Excoffier & Slatkin, 1995; Niu et al., 2002; Kimmel & Shamir, 2004).

This progress notwithstanding, the haplotype models developed so far are still limited in their scope and are inadequate for addressing many realistic problems. Consider for example a genetic demography study, in which one seeks to uncover ethnic- and/or geographic-specific genetic patterns based on a sparse census of multiple populations. In particular, suppose that we are given a sample that can be divided into a set of subpopulations; e.g., African, Asian and European. We may not only want to discover the sets of haplotypes within each subpopulation, but we may also wish to discover which haplotypes are shared between subpopulations, and what are their frequencies. Empirical and theoretical evidence suggests that an early split of an ancestral population following a populational bottleneck (e.g., due to sudden migration or environmental changes) may lead to ethnic-group-specific populational diversity, which features both ancient haplotypes (that have high variability) shared among different ethnic groups, and modern haplotypes (that are more strictly conserved) uniquely present in different ethnic groups (Pritchard, 2001). This structure is analogous to a hierarchical clustering setting in which different groups comprising multiple clusters may share clusters with common centroids (e.g., different news topics may share some common key words).

A naive solution to the aforementioned problem would be to infer haplotypes separately in the subpopulations, using, say, separate DP mixtures. This is clearly suboptimal, however, because it may unnecessarily fragment the data, and may lead to unrobust estimation of demographic parameters. In particular, for rare haplotypes that are present in a small number of individuals (e.g., one or two) in each population but overall still have many bearers across all populations, the estimation of their founders (i.e., the centroid) should take into account of these bearers in all populations jointly, rather than being based on each population separately. Essentially, what we want is a model to solving multiple clustering problems simultaneously. Each clustering can be modeled by an infinite mixture, and the centroids of different clustering problems can be shared. We can formulate this problem using a hierarchical infinite mixture. In a hierarchical infinite mixture, we have a finite number of infinite mixtures, each corresponds to a specific empirically defined population. The components in each of the mixtures can be shared. The hierarchical Dirichlet process (HDP) mixture model developed by Teh et al. (2006) provides a Bayesian approach to capturing exactly such structure.

In this paper, we present **HDP-Haplotyper**, a new statistical genetic model for Bayesian multi-population haplotype inference. This model conjoins a hierarchical Dirichlet process to the haplotype model proposed in Xing et al. (2004), leading to a general Bayesian inference algorithm for jointly inferring haplotypes in multiple populations. Our model is in fact more general than the haplotyping domain. However, apart from its intrinsic interest, this domain has a significant advantage as a development domain for hierarchical Bayesian clustering methods—in the haplotype setting the ground truth, i.e., the true haplotypes, can be obtained if desired via (expensive) sequencing experiments (Patil et al., 2001). Also, pedigree information can be exploited in some cases to obtain ground truth.

2. Background

Elaborating on the notational scheme used in Xing et al. (2004), let $G_i^{(j)} = [G_{i,1}^{(j)}, \ldots, G_{i,T}^{(j)}]$ denote the *genotype* of T contiguous SNPs of individual i from ethnic group j. For diploid organisms such as human, we denote the two alleles of a SNP by 0 and 1; thus each $G_{i,t}^{(j)}$ can take on one of four values: 0 or 1, indicating a homozygous site; 2, indicating a heterozygous site; and '?', indicating missing data. (A generalization to polymorphisms with k-ary alleles is straightforward, but omitted here for simplicity.). A haplotype of individual i from ethnic group j is denoted by $H_{i_e}^{(j)} = [H_{i_e,1}^{(j)}, \ldots, H_{i_e,T}^{(j)}]$, where the sub-subscript

$e \in \{0, 1\}$ denotes the two possible parental origins (i.e., paternal and maternal) of the haplotypes.

In the following, we present a probabilistic model for the generation of haplotypes in multiple populations, and for the generation of genotypes from these haplotypes. We assume that each individual's genotype is formed by drawing two random *templates* from an ancestral pool of founding haplotypes typical of the ethnic population to which the individual belongs, and that these templates are subject to random perturbation according to an *inheritance model* $P_h(H|A)$, where H denotes an individual haplotype and A denote its ancestral template. We further assume that the given noisy observations of the resulting genotypes are related to the true haplotypes via an *genotyping model* $P_g(G|H_0, H_1)$. Since the size of the ancestor pool and its composition are both unknown for each of the populations, we treat them as random variables under a hierarchical Dirichlet process prior. We begin by providing a brief description of both the basic Dirichlet process and hierarchical DP, and subsequently show how this process can be incorporated into a model for multi-population haplotype inference.

2.1. Dirichlet process mixtures

For completeness, we begin with a brief recap of the basic DP mixture model for haplotypes, as used in Xing et al. (2004). As introduced by (Ferguson, 1973), a Dirichlet process (DP) is the distribution of a *random probability distribution* G on some sample space, such that for any partition $(A_1, \ldots, A_k)$ of the sample space, we have:

$$(Q(A_1), \ldots, Q(A_k)) \sim \mathrm{Dir}(\tau Q_0(A_1), \ldots, \tau Q_0(A_k)),$$

where Dir denotes the finite-dimensional Dirichlet distribution, where τ denotes a *concentration parameter* and where Q_0 denotes a *base measure* over the sample space. We write $Q \sim DP(\tau, Q_0)$ if Q is distributed according to a DP.

A more concrete understanding of the DP can be obtained by considering a *Pólya urn model*, a distribution on labeled partitions of data. Consider in particular an urn that at the outset contains a ball of a single color. At each step we either draw a ball from the urn and replace it with two balls of the same color, or we are given a ball of a new color which we place in the urn. The parameter τ defines the probabilities of these two cases. Viewing each (distinct) color as a sample from Q_0 and each ball as a sample from Q, Blackwell and MacQueen (1973) showed that this Pólya urn model yields samples whose distributions are those of the marginal probabilities under the Dirichlet process. The urn scheme also directly suggests a sampling-based computational scheme for posterior inference.

In the context of mixture models, we associate mixture components with colors in the Pólya urn model and thereby define a "clustering" of the data. Specifically, let ϕ_i denote the choice of mixture component associated with the data point x_i; i.e., let $x_i \sim f(\cdot|\phi_i)$ for some likelihood function f. Under the DP model, ϕ_i is a sample from Q. Note that since many balls in a Pólya urn have the same color, and the number of colors is not fixed but grows with the number of data points, we end up with a mixture model for $(x_1, x_2, x_3, \ldots)$ with unbounded cardinality. This mixture model is referred to as a DP mixture. Xing et al. (2004) used this construction to define a generative model for genotypes in a single population as follows:

- Draw first haplotype:

  $\phi_1 \mid \mathrm{DP}(\tau, Q_0) \sim Q_0(\cdot)$, sample the 1st founder;

  $h_1 \sim P_h(\cdot|\phi_1)$, sample the 1st haplotype from an inheritance model defined on the 1st founder;

- for subsequent haplotypes:

  – sample the founder indicator for the ith haplotype:

$$c_i|\mathrm{DP}(\tau, Q_0) \sim \begin{cases} p(c_i = c_j \text{ for some } j < i|c_1, \ldots, c_{i-1}) = \frac{n_{c_j}}{i-1+\alpha_0} \\ p(c_i \neq c_j \text{ for all } j < i|c_1, \ldots, c_{i-1}) = \frac{\alpha_0}{i-1+\alpha_0} \end{cases}$$

  where n_{c_i} is the *occupancy number* of class c_i—the number of previous samples belonging to class c_i.

  – sample the founder of haplotype i (indexed by c_i):

$$\phi_{c_i}|\mathrm{DP}(\tau, Q_0) \begin{cases} = \phi_{c_j} & \text{if } c_i = c_j \text{ for some } j < i \text{ (i.e., } c_i \text{ refers to an inherited founder)} \\ \sim Q_0(\phi) & \text{if } c_i \neq c_j \text{ for all } j < i \text{ (i.e., } c_i \text{ refers to a new founder)} \end{cases}$$

  – sample the haplotype according to its founder:
  $h_i \mid c_i \sim P_h(\cdot|\phi_{c_i})$.

- sample all genotypes (according to a one-to-one mapping between haplotype index i and allele index n_e):

  $g_n \mid h_{n_0}, h_{n_1} \sim P_g(\cdot|h_{n_0}, h_{n_1})$.

2.2. Hierarchical Dirichlet process mixtures

Now we consider the case in which there exist multiple sample populations (e.g., ethnic groups), each modeled by a distinct DP mixture. The components (e.g., ancestors) in any of the mixtures may be shared across the mixtures, but the *weight* of a component in each mixture is unique. To tie population-specific DP mixtures together in this way, we develop a hierarchical DP mixture model (Teh et al., 2006), in which the base measure associated with each population-specific DP mixture is itself drawn from a Dirichlet process $\mathrm{DP}(\gamma, F)$. Since a draw from a DP is a discrete measure with probability 1, atoms drawn from this measure—atoms which are used as the mixture components in each of the population-specific DP mixtures—are not generally distinct. This allows sharing of components across mixture models.

As with the DP, it is useful to describe the marginals induced with an HDP using the more concrete representation of Pólya urn models. Imagine we set up a single "stock" urn at the top level, which contains balls of colors that are represented by at least one ball in one or multiple urns at the bottom level. At the bottom level, we have a set of *distinct* urns which are used to define the DP mixture for each population. Now let's suppose that upon drawing the m_j-th ball for urn j at the bottom, the stock urn contains n balls of K distinct colors indexed by an integer set $\mathcal{C} = \{1, 2, \ldots, K\}$. Now we either draw a ball randomly from urn j, and place back two balls both of that color, or with some probability we return to the top level. From the stock urn, we can either draw a ball randomly and put back two balls of that color in the stock urn and one in j, or obtain a ball of a new color $K+1$ with probability $\frac{\gamma}{n-1+\gamma}$ and put back a ball of this color in both the stock urn and urn j of the lower level. Essentially, we have a master DP (the top urn) that serves as a source of atoms for J child DPs (bottom urns). Our Pólya urn scheme is similar in spirit to the urn scheme discussed in Teh et al. (2006), but it differs in that it avoids having separate occupancy counters in each lower-level DP for repeated draws of the same atom from a top-level DP.

Associating each color k with a random variable ϕ_k whose values are drawn from the base measure F, and recalling our discussion in the previous section, we know that draws from the stock urn can be viewed as marginals from a random measure distributed as a Dirichlet Process Q_0 with parameter (γ, F). Specifically, for n random draws $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_n\}$ from Q_0, the conditional prior for $(\phi_n|\boldsymbol{\phi}_{-n})$, where the subscript "$-n$" denotes the index set of all but the n-th ball, is

$$\phi_n|\boldsymbol{\phi}_{-n} \sim \sum_{k=1}^{K} \frac{n_k}{n-1+\gamma}\delta_{\phi_k^*}(\phi_n) + \frac{\gamma}{n-1+\gamma}F(\phi_i), \quad (1)$$

where $\phi_k^*, k = 1, \ldots, K$ denote the K distinct values (i.e., colors) of $\boldsymbol{\phi}$ (i.e., all the balls in the stock urn), n_k denote the number of balls of color k in the top urn, and $\delta_a(\phi_i)$ denotes a unit point mass at $\phi_i = a$.

Conditioning on Q_0 (i.e., using Q_0 as an atomic base measure of each of the DPs corresponding to the bottom-level urns), the m_j-th draws from the jth bottom-level urn are also distributed as marginals under a Dirichlet measure:

$$\begin{aligned}
&\phi_{m_j}|\boldsymbol{\phi}_{-m_j} \\
&\sim \sum_{k=1}^{K} \frac{m_{j,k} + \tau\frac{n_k}{n-1+\gamma}}{m_j - 1 + \tau}\delta_{\phi_k^*}(\phi_{m_j}) + \frac{\tau}{m_j-1+\tau}\frac{\gamma}{n-1+\gamma}F(\phi_{m_j}) \\
&= \sum_{k=1}^{K} \pi_k \delta_{\phi_k^*}(\phi_{m_j}) + \pi_{K+1}F(\phi_{m_j}), \qquad (2)
\end{aligned}$$

where $\pi_k := \frac{m_{j,k}+\tau\frac{n_k}{n-1+\gamma}}{m_j-1+\tau}$, $\pi_{K+1} = \frac{\tau}{m_j-1+\tau}\frac{\gamma}{n-1+\gamma}$, and $m_{j,k}$ denotes the number of balls of color k in the j-th bottom urn.

3. The HDP-Haplotyper model

Using the HDP construction described in the previous section, we now define an HDP mixture model for the genotypes in J populations. The basic generative structure of the model is as follows:

$Q_0\|\gamma, F \sim \mathrm{DP}(\gamma, F)$,	sample a DP of founders for all populations;
$Q_j\|\tau, Q_0 \sim \mathrm{DP}(\tau, Q_0)$,	sample the DP of founder for each population;
$\phi_{i_e}^{(j)}\|Q_j \sim Q_j$,	sample the founder of haplotype i_e in population j;
$h_{i_e}^{(j)}\|\phi_{i_e}^{(j)} \sim P_h(\cdot\|\phi_{i_e}^{(j)})$,	sample haplotype i_e in population j;
$g_i^{(j)}\|h_{i_0}^{(j)}, h_{i_1}^{(j)} \sim P_g(\cdot\|h_{i_0}^{(j)}, h_{i_1}^{(j)})$,	sample genotype i in population j,

where in practice the first three sampling steps follow the nested Pólya urn scheme described above. Note that in the HDP the base measure of each lower-level DP is a draw from the root $\mathrm{DP}(\gamma, F)$. From this description, it is apparent that the totality of all atomic samples (i.e., ancestors) from the base measure F form the common support of both the root DP and all the population-specific DPs. The child DPs place different mass distributions on this common support.

We now discuss the parameterization of the model in more detail, specifically the inheritance model $P_h(\cdot)$ and the genotyping local $P_g(\cdot)$. We follow the presentation in Xing et al. (2004), making modest extensions where necessary.

Recall that a base measure F at the root of HDP is defined as a distribution from which ancestor haplotype templates ϕ_k are drawn. We define $\phi_k := \{A_k, \theta_k\}$, where $A_k := [A_{k,1}, ..., A_{k,T}]$ is a founding *haplotype configuration* for loci $t = [1, ..., T]$ and define θ_k as the *mutation rate* of this founder. The latter denotes the probability that an allele at a locus is identical to the ancestor at this locus. Under this framework, the base measure F is a joint measure on both A and θ. We let $F(A,\theta) = p(A)p(\theta)$, where $p(A)$ is uniform over all possible haplotypes and $p(\theta)$ is a beta distribution, $Beta(\alpha_h, \beta_h)$, with a small value for $\beta_h/(\alpha_h+\beta_h)$ corresponding to a prior expectation of a low mutation rate. For simplicity, we assume each $A_{k,t}$ (and also each $H_{i,t}^{(j)}$) takes its value from an allele set B. Omitting all but the locus index t, we define our inheritance model to be a *single-locus mutation model* as follows (Xing et al., 2004):

$$p(h_t|a_t,\theta) = \theta^{\mathbb{I}(h_t=a_t)} \left(\frac{1-\theta}{|B|-1}\right)^{\mathbb{I}(h_t\neq a_t)} \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Let J denote the total number of populations (i.e., ethnic groups), let I_j denote the number of individuals in j-th group, and assume that the population identity (i.e., ethnic identity) of each individual is known (although unknown population identity can be addressed by introducing latent population indicator variables). For each individual haplotype $H_{i_e}^{(j)}$, define $C_{i_e}^{(j)}$ to be its founder indicator. With this setup, the joint conditional probability of haplotype instances $\mathbf{h} = \{h_{i_e}^{(j)} : e \in \{0,1\}, i \in \{1,2,...,I_j\}, j \in \{1,2,...,J\}\}$ and the mutation rates $\theta = \{\theta_1,...,\theta_K\}$, given the ancestor indicator $\mathbf{c} = \{c_{i_e}^{(j)} : e \in \{0,1\}, i \in \{1,2,...,I_j\}, j \in \{1,2,...,J\}\}$ of haplotypes and the set of ancestors $\mathbf{a} = \{a_1,...,a_K\}$, can be written explicitly as:

$$p(\mathbf{h},\theta|\mathbf{a},\mathbf{c}) = \prod_{k=1}^{K} \frac{R(\alpha_h,\beta_h)}{(|B|-1)^{l'_k}} [\theta_k]^{\alpha_h+l_k-1}[1-\theta_k]^{\beta_h+l'_k-1} \tag{4}$$

where $R(\alpha_h,\beta_h) = \frac{\Gamma(\alpha_h+\beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)}$, $l_k = \sum_{j,i,e,t}\mathbb{I}(h_{i_e,t}^{(j)} = a_{k,t})\mathbb{I}(c_{i_e}^{(j)} = k)$ is the number of alleles in all populations which are identical to the ancestral alleles, and $l'_k = \sum_{j,i,e,t}\mathbb{I}(h_{i_e}^{(j)} \neq a_{k,t})\mathbb{I}(c_{i_e}^{(j)} = k)$ is the total number of mutated alleles. The marginal conditional distribution of haplotype instances can be obtained by integrating out θ in Eq. (4):

$$\begin{aligned} & p(\mathbf{h}|\mathbf{a},\mathbf{c}) \\ &= \prod_{k=1}^{K} R(\alpha_h,\beta_h) \frac{\Gamma(\alpha_h+l_k)\Gamma(\beta_h+l'_k)}{\Gamma(\alpha_h+\beta_h+l_k+l'_k)} \left(\frac{1}{|B|-1}\right)^{l'_k} \end{aligned} \tag{5}$$

Our genotyping model P_g is the same as in Xing et al. (2004); it assumes that the observed genotype at a locus is determined by the paternal and maternal alleles of this site, subject to low-probability corruptions (i.e., inconsistencies).

For the concentration parameters γ and τ, we use vague inverse Gamma priors:

$$p(\gamma^{-1}) \sim \mathcal{G}(1,1) \Rightarrow p(\gamma) \propto \gamma^{-2}\exp(-1/\gamma)) \tag{6}$$

and similarly for τ. The posterior distribution of γ depends only on the number of instances and the number of classes. The predictive distribution of γ is:

$$\begin{aligned} p(n_1,\ldots,n_k|\gamma) &\propto \frac{\gamma^k\Gamma(\gamma)}{\Gamma(n+\gamma)} \\ p(\gamma|k,n) &\propto \frac{\gamma^{k-2}\exp(1/\gamma)\Gamma(\gamma)}{\Gamma(n+\gamma)}. \end{aligned} \tag{7}$$

The conditional posterior for γ depends only on the number of samples, n, and the number of components, k, and not on how the samples are distributed among the components. The distribution $p(\log(\gamma)|k,n)$ is log-concave, so we may efficiently generate independent samples from this distribution using adaptive rejection sampling (Rasmussen, 2000).

Finally, note that we have used a single concentration parameter τ for the lower-level DPs; it is also possible to allow separate concentration parameters for each of the lower-level DPs, possibly tied distributionally via a common hyperparameter.

4. Inference

In this section we describe a Gibbs sampling algorithm for posterior inference under the HDP-Haplotyper model. The variables of interest are $c_{i_e,t}^{(j)}, a_{k,t}$, $h_{i_e,t}^{(j)}$ and $g_{i,t}^{(j)}$ (the only observed variables). We may assume that the represented mixture components (i.e., founders) are indexed by $1,...,K$, the weights of the founders (i.e., the mixing proportions) at the top level DP is $\beta = (\frac{n_1}{n-1+\gamma},...,\frac{n_K}{n-1+\gamma},\frac{\gamma}{n-1+\gamma})$ where $\frac{\gamma}{n-1+\gamma}$ is the total weight corresponding to some unrepresented founder $K+1$; and the weights of founders at the bottom-level DP for, say, the jth population, are $(\frac{m_{j,1}}{m_j-1+\tau},...,\frac{m_{j,K}}{m_j-1+\tau},\frac{\tau}{m_j-1+\tau})$, where $\frac{\tau}{m_j-1+\tau}$ corresponds to the probability of consulting the top-level DP. The Gibbs sampler alternates between two coupled stages. First, we sample the $c_{i_e}^{(j)}$ and $a_{k,t}$ given the current values of the hidden haplotypes. Then, given the current state of the ancestral pool and the ancestral template assignment for each individual, we sample the $h_{i_e,t}^{(j)}$ variables.

Before sampling $c_{i_e}^{(j)}$, we first erase its contribution to the sufficient statistics of the model. If the old $c_{i_e}^{(j)}$ was k', set $m_{jk'} = m_{jk'} - 1$. If it was sampled from the top level DP, we also set $n_{k'} = n_{k'} - 1$. Note that $c_{i_e}^{(j)} \leq K+1$ (i.e., indicating existing founders, plus a new one to be instantiated). Now we can sample $c_{i_e}^{(j)}$ from the following conditional distribution:

$$\begin{aligned} & p(c_{i_e}^{(j)} = k | \mathbf{c}^{[-j,i_e]}, \mathbf{h}, \mathbf{a}) \\ \propto\ & p(c_{i_e}^{(j)} = k | \mathbf{c}^{[-j,i_e]}, \mathbf{m}, \mathbf{n}) p(h_{i_e}^{(j)} | a_k, \mathbf{c}, \mathbf{h}^{[-j,i_e]}) \\ \propto\ & (m_{jk}^{[-j,i_e]} + \tau\beta_k) p(h_{i_e}^{(j)} | a_k, \mathbf{l}_k^{[-j,i_e]}), \text{ for } k = 1, ..., K+1 \end{aligned}$$

where $m_{jk}^{[-j,i_e]}$ represents the number of $c_{i'_{e'}}^{(j)}$ that are equal to k, except $c_{i_e}^{(j)}$ in group j, and $m_{j,K+1} = 0$; $\mathbf{l}_k^{[-j,i_e]}$ denotes the sufficient statistics associated with all haplotype instances originating from ancestor k, except $h_{i_e}^{(j)}$. If as a result of sampling $c_{i_e}^{(j)}$ a formerly represented founder is left with no haplotype associated with it, we remove it from the represented list of founders. If on the other hand the selected value k is not equal to any other existing index $c_{i_e}^{(j)}$, i.e, $c_{i_e}^{(j)} = K+1$, we increment K by 1, set $n_{K+1} = 1$, update β accordingly, and sample a_{K+1} from its base measure F. We can also use the approximate Metropolis-Hasting updating proposed in Xing et al. (2004) to speed up the mixing of the Markov chain.

Now, from Equation (5), we can use the following posterior distribution to sample the founder a_k:

$$\begin{aligned} p(a_{k,t} | \mathbf{c}, \mathbf{h}) &\propto \prod_{j, i_e | c_{i_e,t}^{(j)} = k} p(h_{i_e,t}^{(j)} | a_{k,t}, l_{k,t}^{(j)}) \\ &= \frac{\Gamma(\alpha_h + l_{k,t}) \Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h + \beta_h + m_k)(|B| - 1)^{l'_{k,t}}} R(\alpha_h, \beta_h) \quad (8) \end{aligned}$$

where $l_{k,t}$ is the number of allelic instances originating from ancestor k at locus t across the groups that are identical to the ancestor, when the ancestor has the pattern $a_{k,t}$. If k was not represented previously, we can just use zero values of $l_{k,t}$ which is equivalent to using the probability $p(a | h_{i_e}^{(j)})$.

We now proceed to the second sampling stage, in which we sample the haplotypes $h_{i_e}^{(j)}$ according to the following conditional distribution:

$$\begin{aligned} & p(h_{i_e,t}^{(j)} | \mathbf{h}_{[-i_e,t]}^{(j)}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \\ & \quad \propto p(g_{i,t}^{(j)} | h_{i_e,t}^{(j)}, h_{i_{\overline{e}},t}^{(j)}, \mathbf{u}_{[-i_e,t]}^{(j)}) p(h_{i_e,t}^{(j)} | a_{k,t}, \mathbf{l}_{[-i_e,t]}^{(j)}) \quad (9) \end{aligned}$$

where $l_{k,t}^{(j)} = l_{k,t}^{(j)[-i_e,t]} + \mathbb{I}(h_{i_e,t}^{(j)} = a_{k,t})$ and $\mathbf{u}_{[-i_e,t]}^{(j)}$ are the set of sufficient statistics recording the inconsistencies between the haplotypes and genotypes in population j. See Eq. (11) in Xing et al. (2004) for detailed definitions and parameterizations.

5. Experiments

To validate the HDP-Haplotyper algorithm, we applied it to both simulated data and real data. We compare its performance to those of DP-Haplotyper (Xing et al., 2004), PHASE 2.1.1 (Stephens et al., 2001; Stephens & Scheet, 2005) and HAPLOTYPER 1.0 (Niu et al., 2002). We ran each program using its default values. Two different error measures are used for evaluation: err_s, the ratio of incorrectly phased SNP sites over all non-trivial heterozygous SNPs (excluding individuals with a single heterozygous SNP), and err_i, the ratio of incorrectly phased individuals over all non-trivial heterogeneous individuals (i.e., those with at least two heterogeneous SNPs). We also present the number of reconstructed founder haplotypes K (note that HAPLOTYPER and PHASE do not infer the number of ancestors and thus the K there merely means the total number of distinct haplotypes).

5.1. Simulated data

To simulate multi-population genotypes/haplotypes, we collected a candidate pool of founding haplotypes from the dataset used in Stephens et al. (2001). Among them, a few haplotype templates are selected as shared founders across different groups, and for each group, more templates are added to form group-specific pools of founders. We then drew each individual's genotype and haplotypes by randomly choosing two templates from the pool and applying the mutation and noisy observation process described in Section 3.

The synthetic dataset includes 100 individuals from five groups (20 from each), with genotypes containing 10 sites. Each group has two shared founders and three more templates unique to the group. So, overall each group has 5 founders, while the total number of founders across the five groups is 17. We tested on two datasets with different mutation rates, 0.01 and 0.05, respectively. The noisy observation rate was the same for each group, each individual, and each locus.

In our first experiment, we type all 100 individuals together. Our HDP approach makes use of the group label information, while the other algorithms (DP, PHASE and HAPLOTYPER) ignore such information, treating all individuals as if they are from a single population. Table 1 summarizes the performance of each algorithm. We see that HDP outperformed the other algorithms on both datasets. Note that we expect K to be 17, and the MAP estimates under the DP and HDP models turn out to be very close to this number (we omit a plot of the full posterior due to space limitations); note also that the parametric methods (PHASE and HAPLOTYPER) can not provide an estimate of this quantity.

Next, to see the effectiveness of simultaneous multiple clustering via the HDP mixture, we compared it

Table 1. Performance on a simulated dataset. Two kinds of datasets with different mutation rates θ were tested. Each dataset includes 100 individuals from 5 groups (20 from each). The sequence length was fixed to 10. The performance of each algorithm is represented in terms of err_s, err_i and K.

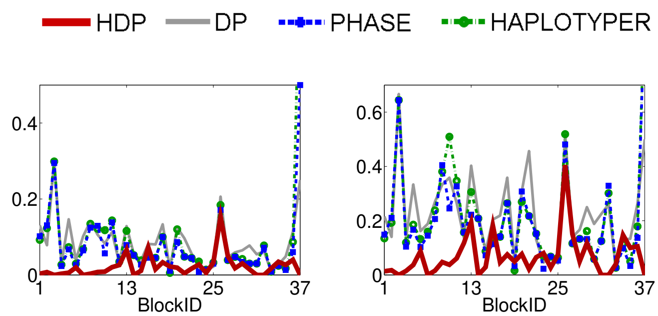
		HDP			DP			PHASE			HAPLOTYPER		
θ	I	err_s	err_i	K	err_s	err_i	K	err_s	err_i	K	err_s	err_i	K
0.01	100	0.0133	0.0589	15	0.0229	0.0732	17	0.0287	0.0976	36	0.0350	0.0976	34
0.05	100	0.1076	0.3626	17	0.1777	0.4944	24	0.1920	0.5840	73	0.2006	0.5618	54

to results from separate runs of the other algorithms on each group of genotypes (Table 2). The group-wise results from the HDP were extracted from the runs shown previously in Table 1. Here, K is expected to be 5 for each group, and HDP again yields a MAP estimate that is close to this value. On the dataset with mutation rate 0.01, all algorithms performed comparably. On the dataset with a higher mutation rate (0.05), the HDP approach outperformed other algorithms and inferred the founders of each group more robustly than the group-specific runs based on separate DP mixtures.

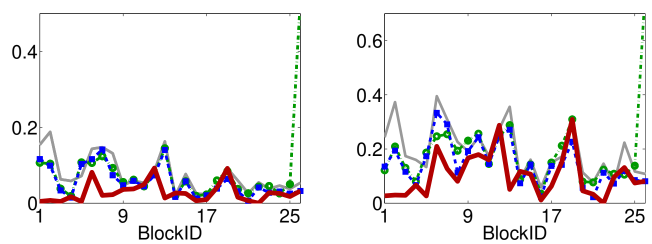
5.2. Real data

We applied our algorithm to the database from the International HapMap Project, which contains four populations of genotypes. Among them, we focused on two populations of CEPH (Utah residents with ancestry from northern and western Europe, CEU) and Yoruba in Ibadan, Nigeria (YRI) since they contain 30 trios of genotypes and allow us to infer most of the true haplotypes. The common SNP sites of length 254 could be extracted from the region *ENm010.7p15.2*.

For computational and biological reasons (e.g., the presence of recombination), we partitioned the SNP sites into blocks of shorter lengths; following a recommendation in Niu et al. (2002), we used block lengths no greater than 10. The results for block length 7 and 10 are shown in Fig. 1. For block length 7, the average err_s across the 37 blocks was 0.0228, 0.0864, 0.0760, and 0.0928 for HDP, DP, PHASE, and HAPLOTYPER, respectively; and the average err_i was 0.0689, 0.2325, 0.2002, and 0.2158, respectively, in the same order. We thus see significant performance improvements under the HDP model. When tested with block length 10, the average err_s across the 26 blocks were 0.0276, 0.0754, 0.0565, and 0.0813, and the average err_i across the blocks were 0.1024, 0.1946, 0.1467, and 0.1901, respectively.



(a) Performance on 37 blocks with length 7



(b) Performance on 26 blocks with length 10

Figure 1. Performance on HapMap data, with 254 SNPs partitioned into (a) 37 blocks of length 7 and (b) 26 blocks of length 10. The left panels represent err_s for each block and the right panels represent err_i.

6. Conclusions

We have proposed a new Bayesian approach to haplotype inference for multiple populations using a hierarchical Dirichlet process mixture. By incorporating an HDP prior which couples multiple heterogeneous populations and facilitates sharing of mixture components across multiple infinite mixtures, the proposed method can infer the true haplotypes in a multi-ethnic group with an accuracy superior to state-of-the-art haplotype inference algorithms.

The experiments presented in this paper focus on relatively short sequences of SNPs. In ongoing work, we are developing a Partion-Ligation scheme based on the work in Niu et al. (2002) to deal with longer sequences.

Finally, although in the present study we have assumed that the population structure—the ethnic labels of individuals—are known, it is straightforward to generalize our method to situations in which the ethnic group labels are unknown and to be inferred. This opens the door to applications of our method

Table 2. Comparison of HDP and other algorithms on the multi-population data. The results for HDP were extracted from the corresponding run in Table 1. The results for the other algorithms were obtained by running the algorithms separately on the different groups and averaging.

			HDP			DP			PHASE			HAPLOTYPER		
θ	I	group	err_s	err_i	K	err_s	err_i	K	err_s	err_i	K	err_s	err_i	K
0.01	20	(1)	0.0159	0.0556	5	0.0159	0.0556	5	0.0000	0.0000	9	0.0159	0.0556	8
	20	(2)	0.0000	0.0000	5	0.0175	0.0590	5	0.0000	0.0000	7	0.0526	0.0588	6
	20	(3)	0.0141	0.0625	4	0.0000	0.0000	5	0.0000	0.0000	9	0.0000	0.0000	8
	20	(4)	0.0366	0.1765	4	0.0244	0.0590	5	0.0366	0.1765	9	0.0244	0.1176	8
	20	(5)	0.0000	0.0000	5	0.0244	0.0710	7	0.0488	0.0714	11	0.0732	0.1429	10
		avg	0.0133	0.0589		0.0164	0.0489		0.0171	0.0496		0.0332	0.0749	
0.05	20	(1)	0.0758	0.2780	5	0.0758	0.3330	6	0.1970	0.6111	20	0.0758	0.2222	14
	20	(2)	0.1640	0.5000	5	0.1640	0.5560	8	0.1148	0.3333	17	0.1967	0.4444	15
	20	(3)	0.0886	0.4120	5	0.1140	0.5290	5	0.1013	0.4706	17	0.1139	0.5294	15
	20	(4)	0.0455	0.2110	5	0.0568	0.3680	10	0.1705	0.6316	22	0.1136	0.4737	15
	20	(5)	0.1640	0.4120	7	0.2180	0.4120	6	0.1818	0.4706	16	0.1273	0.4118	14
		avg	0.1076	0.3626		0.1257	0.4396		0.1531	0.5034		0.1255	0.4163	

to large-scale genetic studies involving joint inference over markers and demography.

Acknowledgments

This research was supported in part by NSF EMT Grant CCF-0523757.

References

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, *1*, 353–355.

Chakravarti, A. (2001). Single nucleotide polymorphisms: . . .to a future of genetic medicine. *Nature*, *409*, 822–823.

Clark, A. (1990). Inferences of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol*, *7*, 111–122.

Clark, A. (2003). Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr Opin Genet Dev*, *13*, 296–302.

Excoffier, L., & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, *12*, 921–7.

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

Group, T. I. S. M. W. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*, 928 – 933.

Hodge, S. E., Boehnke, M., & Spence, M. A. (1999). Loss of information due to ambiguous haplotyping of SNPs. *Nat Genet*, *21*, 360–361.

Kimmel, G., & Shamir, R. (2004). Maximum likelihood resolution of multi-block genotypes. *Proceedings of RECOMB 2004* (pp. 847–56).

Niu, T., Qin, S., Xu, X., & Liu, J. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *American Journal of Human Genetics*, *70*, 157–169.

Patil, N., Berno, A. J., et al. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, *294*, 1719–1723.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex disease? *Am J Hum Genet*, *69*, 124–137.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press.

Stephens, M., & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *American Journal of Human Genetics*, *76*, 449–462.

Stephens, M., Smith, N., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, *68*, 978–989.

Teh, Y., Jordan, M. I., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* (to appear).

Xing, E., Sharan, R., & Jordan, M. (2004). Bayesian haplotype inference via the Dirichlet process. *Proceedings of the 21st International Conference on Machine Learning*. New York, NY: ACM Press.