Bayesian inference on population structure: from parametric to nonparametric modeling

Maria De Iorio, Stefano Favaro and Yee Whye Teh

Abstract Making inference on population structure from genotype data requires to identify the actual subpopulations and assign individuals to these populations. The source populations are assumed to be in Hardy-Weinberg equilibrium, but the allelic frequencies of these populations and even the number of populations present in a sample are unknown. In this chapter we present a review of some Bayesian parametric and nonparametric models for making inference on population structure, with emphasis on model-based clustering methods. Our aim is to show how recent developments in Bayesian nonparametrics have been usefully exploited in order to introduce natural nonparametric counterparts of some of the most celebrated parametric approaches for inferring population structure. We use data from the 1000 Genomes project (http://www.1000genomes.org/) to provide a brief illustration of some of these nonparametric approaches.

1 Introduction

Population stratification or structure refers to the presence of a systematic difference in allele frequencies between populations due to the fact that populations are typically heterogeneous in terms of their genetic ancestry. A particular type of population structure is genetic admixtures, which derive from the genetic mixing of two or more previously separated groups in the recent past. A typical example is offered by African-Americans. The analysis of population structure based on genotypes at

Maria De Iorio
Department of Statistical Science, University College London, UK e-mail: m.deiorio@ucl.ac.uk

Stefano Favaro
Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto, Italy e-mail: stefano.favaro@unito.it

Yee Whye Teh
Department of Statistics, University of Oxford, UK e-mail: y.w.teh@stats.ox.ac.uk

co-dominant marker loci presents an important problem in population genetics. In particular it is central to the understanding of human migratory history and the genesis of modern populations, while the associated admixture analysis of individuals is important in correcting the confounding effects of population ancestry in gene mapping and association studies. As allele frequencies are known to vary among populations of different genetic ancestry, similarly phenotypic variation, such as disease risk, is observed among group of different genetic ancestry. Population structure is also relevant in the analysis of gene flow in hybridization zones (Field et al., 2011) and invasive species (Ray and Quader, 2014), conservation genetics (Wasser et al., 2007) and domestication events (Park et al., 2004).

The advent of high density genotyping arrays and next generation resequencing technologies has led to the production of enormous quantity of data, offering an opportunity to investigate ancestry and genetic relationships among individuals in a population in unprecedented level of details. Nevertheless, this enormous quantity of available data poses new statistical and computational challenges. Making inference on population structure from genotype data requires to identify the actual subpopulations and, in particular, assign individuals to these populations. The source populations are assumed to be in the Hardy-Weinberg equilibrium, namely the likelihood of the genotype of an individual, conditional on its subpopulation of origin, is simply the product of the frequencies of its alleles in that population. The allelic composition of these populations and even the number of populations are unknown and, therefore, object of inference.

A full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, have been proposed for inferring population structure. Two of the prevailing approaches used to infer genetic ancestry from a sample of chromosomes are Principal Components Analysis (PCA) and Structured Association. PCA has been used to infer population structure from genetic data for several decades (Novembre and Stephens, 2008) and consists in projecting individuals in a lower dimensional space so that the locations of individuals in the projected space reflect the genetic similarities among them. Clusters of individuals in the projected space can be interpreted as genetic populations, while admixture of two populations results in sets of individuals lying along a line. PCA is a computationally efficient method which can handle large numbers of markers, and is useful for visualizing population structure. The first few principal components are often used to correct for population stratification in genetic association studies. PCA is implemented in EIGENSTRAT (Patterson et al., 2006). In a structured association approach the goal is to explicitly infer genetic ancestry: individuals are assigned to subpopulation clusters, possibly allowing fractional cluster membership in the case of genetic admixtures. Techniques from model based clustering are usually employed.

In this chapter we focus on structured association methods, in particular concentrating on Bayesian approaches which model population structure and admixture using mixture models. An influential early Bayesian parametric mixture model has been proposed by Pritchard et al. (2000). Specifically, assuming that marker loci are unlinked and at linkage equilibrium with one another within populations, each individual is assumed to come from one of K populations and alleles at dif-

ferent loci are modeled conditionally independently given population specific allele frequencies. In the case of genetic admixtures, each individual is associated with proportions of its genome coming from different populations, while alleles at different loci are suitably modeled conditionally independently given the admixture proportions. Independent prior distributions on the allelic profile parameters of each population are introduced and full posterior inference is performed through Markov chain Monte Carlo (MCMC) algorithms. With regards to the determination of the number of extant populations K, Pritchard et al. (2000) proposed the use of model selection techniques based on marginal likelihoods, though it has been noted that such estimates are highly sensitive to the prior specification.

Falush et al. (2003) improved the admixture model of Pritchard et al. (2000) by taking into account the correlations among neighboring loci. In particular Falush et al. (2003) model linked loci by using a Markov model which segments each chromosome into contiguous regions with shared genetic ancestry. This Markov model allows for the estimation of local genetic ancestry information from genotype data, as opposed to the global admixture proportions in Pritchard et al. (2000). Such local ancestry estimation gives more fine-grained information about the admixture process. The nonparametric counterpart of the simple population structure model in Pritchard et al. (2000) is described in Huelsenbeck and Andolfatto (2007), while the Hierarchical Dirichlet process of Teh et al. (2006) offers the Bayesian nonparametric extension of the admixture model. Recently, De Iorio et al. (2015) have proposed a Bayesian nonparametric counterpart of the linkage admixture model of Falush et al. (2003). In particular the nonparametric approach provides a methodology for modeling population structure that simultaneously gives estimates of local ancestries and bypasses difficult model selection issues arising in the parametric models by Pritchard et al. (2000) and Falush et al. (2003).

The chapter is structured as follows. In Section 2 we review the Bayesian parametric approaches introduced by Pritchard et al. (2000) and Falush et al. (2003) for modeling population structure with and without admixture and in presence of linked loci and correlated allele frequencies. In Section 3 we show how recent developments in Bayesian nonparametrics have been usefully exploited in order to introduce natural nonparametric counterparts of the parametric approaches by Pritchard et al. (2000) and Falush et al. (2003). Some of these Bayesian nonparametric approaches are briefly illustrated using data from the 1000 Genomes project (http://www.1000genomes.org/). The goal of the 1000 Genomes project consists in finding most genetic variants that have frequencies of at least 1% in the populations under study by sequencing the genomes of a large number of individuals, providing in this way a valuable resource on human genetic variation.

2 Parametric Modeling

Suppose we sample N haploid individuals at L loci from a population with unknown structure. For simplicity we discuss the haploid case, extension to the diploid case

is straightforward. We denote by $X = (X_l^{(i)})_{1\leq i\leq N, 1\leq l\leq L}$ the observed data, i.e., $x_l^{(i)}$ is the genotype of individual i at locus l. Assuming K subpopulations characterized by a set of allele frequencies at each locus, and with K being fixed, in this section we review some Bayesian parametric models for making inference on the unknown population structure.

2.1 Models with and without admixture

We start by assuming that marker loci are unlinked and at linkage equilibrium with one another within populations. Let $Z = (z^{(i)})_{1\leq i\leq N}$ denote the unknown allocation vector which assigns each individual to a population of origin, i.e., $z^{(i)}$ denotes the population from which individual i originated. Let $Q = (q_k)_{1\leq k\leq K}$ denote the unknown population proportions, i.e., q_k is the proportion of individuals that originated from population k. Furthermore, let J_l be the number of distinct alleles observed at locus l, and let $P = (p_{klj})_{1\leq k\leq K, 1\leq l\leq L, 1\leq j\leq J_l}$ be the unknown allele frequencies in the populations, i.e., p_{klj} is the frequency of allele j at locus l in population k. Throughout this chapter we use "allele copies" to refer to an allele carried at a particular locus by a particular individual.

Under this framework Pritchard et al. (2000) introduced a model without admixture among populations, namely the genome of each individual is assumed to be originated entirely from one of the K populations. Given the population of origin of each individual, the genotype is generated by drawing alleles copies independently from the appropriate population frequency distribution. Formally, the model without admixture is specified as

$$\Pr[z^{(i)} = k \,|\, Q] = q_k \tag{1}$$

and

$$\Pr[x_l^{(i)} = j \,|\, Z, P] = p_{z^{(i)}lj} \tag{2}$$

independently for each $x_l^{(i)}$. This model can be easily extended to diploid or, in general, to polyploid data. For polyploid data the allocation variables $z^{(i)}$'s along each of the chromosomes of individual i form independent vectors. We refer to Falush et al. (2003) for details.

The model (1)-(2) is completed by specifying a prior distribution for Q and P. As regard to Q, Pritchard et al. (2000) assumed that the probability that individual i originated in population k is the same for all k. Hence, they proposed to use the uniform distribution $q_k = 1/K$, independently for all individuals. Different distributions for Q have been considered in Anderson and Thompson (2002) to model cases with some populations being more represented in the sample than others. As regard to P, Pritchard et al. (2000) followed Balding and Nichols (1995) and Ranalla and Mountain (1997) in using the Dirichlet distribution to model the allele frequencies at each locus within each populations, i.e.,

$$p_{kl\cdot} \sim \text{Dirichlet}(\lambda_1, \ldots, \lambda_{J_l}), \tag{3}$$

for the allele frequencies $p_{kl\cdot}$, independently for any k and l. Furthermore, they assumed $\lambda_i = 1$ for all $j = 1,\ldots,J_l$, which gives a uniform distribution over the allele frequencies. By means of (2) and (3) we can use the following MCMC scheme to construct a Markov chain with stationary distribution $\Pr[Z,P\,|\,X]$. Start with initial values $Z^{(0)}$ for Z and, for $m \geq 1$: i) sample $P^{(m)}$ from $\Pr[P\,|\,X,Z^{(m-1)}]$ and ii) sample $Z^{(m)}$ from $\Pr[Z\,|\,X,P^{(m)}]$. For sufficiently large m and c, $(Z^{(m)},P^{(m)})$, $(Z^{(m+c)},P^{(m+c)})$, $(Z^{(m+2c)},P^{(m+2c)})$, ... are approximately random samples from the target distribution $\Pr[Z,P\,|\,X]$.

An obvious limitation of the model without admixture is that, in practice, individuals may have recent ancestors in more than one population. In order to overcome this fundamental drawback, Pritchard et al. (2000) introduced a more flexible model in which only a fraction of the individual's genome is assumed to have originated from one of the K populations. This more general model allows individuals to have mixed ancestry. Let $Z = (z_l^{(i)})_{1\leq i\leq N,1\leq l\leq L}$ be the unknown populations of origin of the allele copies, i.e., $z_l^{(i)}$ is the population from which the allele copies at locus l of individual i is originated. Furthermore, let $Q = (q_k^{(i)})_{1\leq i\leq N,\leq k\leq K}$ be the unknown admixture individual proportions, i.e., $q_k^{(i)}$ is the proportion of the genome of individual i that originated from population k.

Under this more general framework, Pritchard et al. (2000) introduced a model which allows for admixture: given the population of origin of each allele copies, the genotype is generated by drawing alleles copies independently from the appropriate population frequency distribution. Formally, the model with admixture is specified as follows

$$\Pr[z_l^{(i)} = k\,|\,Q] = q_k^{(i)} \tag{4}$$

and

$$\Pr[x_l^{(i)} = j\,|\,Z,P] = p_{z_l^{(i)}lj} \tag{5}$$

independently for each $x_l^{(i)}$. This model can be easily extended to diploid or, in general, to polyploid data. For polyploid data the allocation variables $z^{(i)}$'s along each of the chromosomes of individual i form independent vectors. We refer to Falush et al. (2003) for details.

The admixture model (4)-(5) is completed by specifying a prior distribution for Q and P. Pritchard et al. (2000) proposed the use of the Dirichlet distribution (3) for P, as for the model without admixture. The specification of a prior distribution for Q depends on the type and amount of mixed ancestry we expect to see. In particular Pritchard et al. (2000) proposed the use of a symmetric Dirichlet distribution to model the admixture proportions of each individual. Specifically, they specified the distribution

$$q^{(i)} \sim \text{Dirichlet}(\alpha,\ldots,\alpha) \tag{6}$$

for the admixture proportions $q^{(i)}$, independently for each individual. If α tends to 0 then the admixture model reduces to the model without admixture. Different distributions for Q have been considered in Anderson and Thompson (2002). The following MCMC scheme may be used to sample from $\Pr[Z,P,Q\,|\,X]$. Start

with initial values $Z^{(0)}$ for Z and, for $m \geq 1$: i) sample $P^{(m)}$ and $Q^{(m)}$ from $\Pr[P,Q\,|\,X,Z^{(m-1)}]$, ii) sample $Z^{(m)}$ from $\Pr[Z\,|\,X,P^{(m)},Q^{(m)}]$ and update α using a Metropolis-Hastings step. As before, for sufficiently large m and c, note that $(Z^{(m)},P^{(m)},Q^{(m)})$, $(Z^{(m+c)},P^{(m+c)},Q^{(m+c)})$, $(Z^{(m+2c)},P^{(m+2c)},Q^{(m+2c)})$, ... are approximately random samples from the target distribution $\Pr[Z,P,Q\,|\,X]$.

2.2 Extensions: linked loci and correlated allele frequencies

Falush et al. (2003) extended the admixture model of Pritchard et al. (2000) to allow for linkage between loci. In particular they considered the correlations in ancestry, which cause linkage disequilibrium between linked loci. This linkage disequilibrium naturally occurs because the chromosome is composed of a set of chunks that are derived, as intact units, from one or another of the ancestral populations. In order to model linked loci, Falush et al. (2003) assumed that the breakpoints between successive segments occur as a Poisson process at a rate r per unit of genetic distance, and that the population of origin of each chunk in individual i is independently drawn according to the vector $q^{(i)}$, which continues to represent the admixture proportions of the i-th individual.

More formally the linkage admixture model of Falush et al. (2003) assumes that for each individual i the random variables $z_l^{(i)}$'s are dependent across l and, in particular, they form a reversible Markov chain. Specifically, for any positive r, one has the following specification

$$\Pr[z_1^{(i)} = k\,|\,Q] = q_k^{(i)} \tag{7}$$

and

$$\Pr[z_{l+1}^{(i)} = k'\,|\,z_l^{(i)} = k, Q] = \begin{cases} \mathrm{e}^{-d_l r} + (1-\mathrm{e}^{-d_l r}) q_{k'}^{(i)} & \text{if } k' = k \\ (1-\mathrm{e}^{-d_l r}) q_k^{(i)} & \text{if } k \neq k' \end{cases} \tag{8}$$

independently for each individual, where d_l denotes the genetic distance from locus l to locus $l+1$, assumed known. The admixture model (4)-(5) is recovered by letting $r \to +\infty$. We refer to Falush et al. (2003) for details on the MCMC scheme for sampling from $\Pr[Z,P,Q\,|\,X]$.

Falush et al. (2003) also introduce an extension of the admixture model of Pritchard et al. (2000) in order to allow for correlated allele frequencies, namely the allele frequencies in one population provide information about the allele frequencies in another population. Indeed it is expected that allele frequencies in closely related populations tend to be very similar. In order to model closely related populations, Pritchard et al. (2000) replaced the prior distribution (3) with $p_{kl\cdot} \sim \text{Dirichlet}(f^{(l)}J_l\mu_1^{(l)},\ldots,f^{(l)}J_l\mu_{J_l}^{(l)})$, where $\mu_j^{(l)}$ is the mean sample frequency at locus l, and $f^{(l)} > 0$ determines the strength of the correlations across populations

at locus l. Clearly, when $f^{(l)}$ is large, the allele frequencies in all populations tend to be similar to the mean allele frequencies in the sample.

Alternatively, Falush et al. (2003) assume that the populations all diverged from a common ancestral population at the same time, but allow that the populations may have experienced different amounts of drift since the divergence event. Specifically, let p_{Alj} be the frequency of allele j at locus l in a hypothetical ancestral population A. The K populations in the sample have each undergone independent drift away from the ancestral allele frequencies, at rates parameterized by $F_1,\ldots,F_K$, respectively. More formally,

$$p_{Al\cdot} \sim \text{Dirichlet}(\lambda_1,\ldots,\lambda_{J_l}) \tag{9}$$

independently for each l. Note that the prior distribution has the same form as that used in the model with uncorrelated population frequencies. Then, conditionally on P_A,

$$p_{kl\cdot} \sim \text{Dirichlet}\left(p_{Al1}\frac{1-F_k}{F_k},\ldots,p_{AlJ_l}\frac{1-F_k}{F_k}\right) \tag{10}$$

independently for each population k and for each locus l. According to (10) the size of the parameter F_k tells us about the effective size of population k during the time since divergence, with large values of F_k indicating a smaller effective population size. We refer to Falush et al. (2003) for details on the MCMC scheme for sampling from $\Pr[Z,P,Q\,|\,X]$.

The Bayesian parametric approaches described in this section are implemented in STRUCTURE (http://pritchardlab.stanford.edu/structure.html), which is arguably the most widely used software for estimating genetic ancestry. The reader is referred to Pritchard et al. (2000) and Falush et al. (2007) for a description of the basic algorithms. Extensions can be found in Falush et al. (2007) and Hubisz et al. (2009). ADMIXTURE (http://genetics.ucla.edu/software/admixture/index.html) is an alternative software which provides a faster implementation of a similar model to the one defined in STRUCTURE. In particular ADMIXTURE uses maximum likelihood inference to estimate population allele frequencies and ancestry proportions, rather than sampling from posterior distribution through MCMC algorithms. See, e.g., Alexander et al. (2009) for details.

3 Nonparametric Modeling

The Bayesian parametric models reviewed in Section 2 assume the number of populations K to be fixed. In order to deal with an unknown K, Pritchard et al. (2000) suggest a method based upon an *ad hoc* approximation of the marginal likelihood to determine the number of populations needed to explain the observations. In particular STRUCTURE is run for different values of K, and the number of populations is determined by the value of K which maximises the marginal likelihood of the data. Alternatively, ADMIXTURE uses a cross validation approach to estimate K, by fitting the model on a subset of genotype data and then predicting the excluded geno-

types. Other parametric approaches have been proposed by Corander et al. (2003), Corander et al. (2004) and Evanno et al. (2005). In this section we review some Bayesian nonparametric models for making inference on population structure. In the nonparametric framework both the allocation vectors Z and the number of ancestral populations K are unknown.

3.1 Models with and without admixture

A Bayesian nonparametric counterpart of the model without admixture of Pritchard et al. (2000) has been proposed in Huelsenbeck and Andolfatto (2007). This model makes use of the Dirichlet process by Ferguson (1973), which allows both the assignment of individuals to populations and the number K of populations to be random variables. A simple and intuitive definition of the Dirichlet process follows from the stick-breaking construction introduced by Sethuraman (1994). Specifically, let $(v_j)_{j\geq 1}$ be a collection of independent Beta random variables with parameter $(1,\alpha_0)$, and let $(\theta_i)_{i\geq 1}$ be a collection of random variables, independent of $(v_j)_{j\geq 1}$, and independent and identically distributed according to a nonatomic probability measure G_0. The discrete random probability measure $Q_0 = \sum_{j\geq 1} q_j \delta_{\theta_j}$, with $q_j = v_j \prod_{1\leq l\leq j-1}(1-v_l)$, is a Dirichlet process with parameter $\alpha_0 G_0$.

Here and in the following discussion we denote with $\mathrm{DP}(\alpha_0, G_0)$ the distribution of a Dirichlet process with parameter $(\alpha_0 G_0)$. The Bayesian nonparametric model without admixture introduced by Huelsenbeck and Andolfatto (2007) can be specified as follows

$$\begin{aligned} z^{(i)} \,|\, Q_0 &\overset{\text{iid}}{\sim} Q_0 \\ Q_0 &\sim \mathrm{DP}(\alpha_0, G_0) \end{aligned} \tag{11}$$

and

$$\begin{aligned} x_l^{(i)} \,|\, Z, P &\overset{\text{ind}}{\sim} P_Z \\ P_Z &\sim \mathrm{Dirichlet}(\lambda_1, \ldots, \lambda_{J_l}), \end{aligned} \tag{12}$$

for $i = 1,\ldots,N$ and $l = 1,\ldots,L$. See, e.g., Dawson and Belkhir (2001) and Pella and Masuda (2006) for alternative Bayesian nonparametric models which exploit the Dirichlet process at the allocation level.

The sample Z from Q_0 induces a random partition of $\{1,\ldots,N\}$ which determines the allocation of individuals into a random number K of populations with random frequencies $(n_1,\ldots,n_K)$. The parameter α_0 determines the degree to which individuals are grouped together into the same population. Indeed, Blackwell and MacQueen (1973) show that

$$\Pr[z^{(N)} \in \cdot \,|\, z^{(1)}, \ldots, z^{(N-1)}] = \sum_{i=1}^{K} \frac{n_i}{N-1+\alpha_0} \delta_{\theta_i}(\cdot) + \frac{\alpha_0}{N-1+\alpha_0} G_0(\cdot). \qquad (13)$$

The allocation directed by the predictive distribution (13) can be intuitively described by means of the following Chinese restaurant metaphor. See, e.g., Aldous (1985) for a detailed account. Consider a Chinese restaurant with an unbounded number of tables. Each $z^{(i)}$ corresponds to a customer who enters the restaurant, whereas the distinct values θ_j's correspond to the tables at which the customers sit. Customer i sits at the table indexed by θ_j with probability proportional to the number n_j of customers already seated there, in which case we set $z^{(i)} = \theta_j$, and it sits at a new table with probability proportional to α_0, in which case we increment K by 1, draw θ_K from G_0 and set $z^{(i)} = \theta_K$.

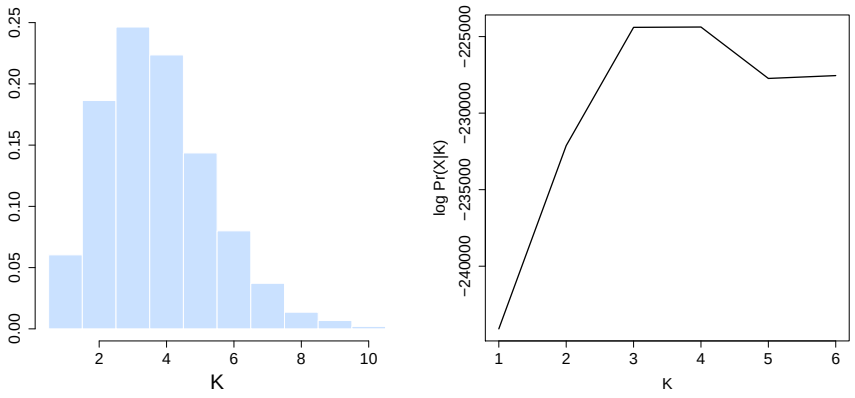

Fig. 1 Left panel: posterior distribution of the number K of populations present in the sample. Right panel: $\log \Pr(\text{Data} \,|\, K)$ from STRUCTURE.

The approach of Huelsenbeck and Andolfatto (2007) is implemented in STRUCTURAMA (http://cteg.berkeley.edu/structurama/). Posterior inference is performed through an MCMC scheme which aims at determining the mean partition, a partitioning of individuals among populations which minimizes the squared distance to the sampled partitions. To illustrate the model (11)-(12), we consider 305 individuals from the 1000 Genomes project. The sample is composed of 95 chromosomes with European ancestry (CEU), 107 chromosomes of African (YRI) origin and 103 individuals of East Asian (CHB) ancestry. In order to phase the genotype data we use SHAPEIT2 (http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html), providing a sample of 610 haplotypes. We have analyzed a collection of 1000 bi-allelic loci from chromosome 2. Posterior inference on the number of populations in the

sample is shown in the left panel of Figure 1. The posterior distribution of K has its mode in 3, which is the true number of populations present in the sample. We have also run the model without admixture implemented in STRUCTURE for each value of K, $K = 1,\ldots,6$. The right panel of Figure 1 shows the estimated $\log \Pr(\text{Data}\,|\,K)$. Note that the value $K = 3,4$ seems to maximizes the model log-likelihood. With this regards, it is worth pointing out that Pritchard et al. (2000) warn against possible drawbacks of using this criterion and to interpret the results with caution and give suggestions for improvement.

A Bayesian nonparametric counterpart of the linkage admixture model of Falush et al. (2003) has been recently introduced and investigated in De Iorio et al. (2015). This model extends the hierarchical Dirichlet process introduced by Teh et al. (2006). The hierarchical Dirichlet process is defined as a distribution over a collection of discrete random probability measures. Specifically, let α_0 and α be positive constants and let G_0 be a nonatomic probability measure. The hierarchical Dirichlet process defines a set of local discrete random probability measures $(Q_i)_{i\in\{1,\ldots,N\}}$, for some index $N \geq 1$, and a global discrete random probability measure Q_0 such that Q_0 is a Dirichlet process with parameter $\alpha_0 G_0$ and, given Q_0, $(Q_i)_{i\in\{1,\ldots,N\}}$ is a collection of independent Dirichlet processes, each one with the same parameter αQ_0. Because the global Q_0 has support at the points $(\theta_i)_{i\geq 1}$, each local random probability measure Q_i necessarily has support at these points as well, and thus can be written as $Q_i = \sum_{j\geq 1} q_{ij}\delta_{\theta_j}$, with $q_{ij} = w_{ij}\prod_{1\leq l\leq j-1}(1-w_{il})$, where $(w_{ij}\,|\,v_1,\ldots,v_j)_{j\geq 1}$ are independent random variables from a Beta distribution with parameter $(\alpha v_j, \alpha(1-\sum_{1\leq l\leq j} v_l))$. Note that the Dirichlet process with parameter $\alpha_0 G_0$ is recovered by letting $\alpha \to 0$.

Due to the sharing of atoms among discrete random random probability measures, the hierarchical Dirichlet process is the natural generalization of the Dirichlet process to model linked sets of admixture proportions and constitutes the Bayesian nonparametric counterpart of the admixture model defined in (4)-(5). Individual genotypes will have portions that arise from different populations which are shared among individuals. The hierarchical Dirichlet process models the allocation vector $Z = (z_l^{(i)})_{1\leq i\leq N, 1\leq l\leq L}$ that specifies the populations of origin of the allele copies. Accordingly, the resulting Bayesian nonparametric "admixture" model can be specified as follows

$$\begin{aligned} z_l^{(i)}\,|\,Q_i &\overset{\text{iid}}{\sim} Q_i \\ Q_i\,|\,Q_0 &\overset{\text{iid}}{\sim} \text{DP}(\alpha, Q_0), \\ Q_0 &\sim \text{DP}(\alpha_0, G_0) \end{aligned} \tag{14}$$

and

$$x_l^{(i)}\,|\,Z,P \overset{\text{ind}}{\sim} P_Z$$

$$P_Z \quad \sim \quad \text{Dirichlet}(\lambda_1, \ldots, \lambda_{J_l}), \tag{15}$$

for $i = 1, \ldots, N$ and $l = 1, \ldots, L$. Note that the hierarchical Dirichlet process prior assumption allows to model Z in terms of an unknown number of populations K with unknown frequencies that are shared out across individuals; each individual's genome is then modeled according to an unknown number of populations with unknown allocation frequencies. For polyploid data the z's along each of the chromosomes of individual i are assumed to be independent.

The allocation mechanism induced by the hierarchical Dirichlet process can be intuitively described by the following generalization of the Chinese restaurant metaphor. Consider a finite collection of Chinese restaurants, one for each index $i \in \{1, \ldots, N\}$, with a shared menu. Each $z_l^{(i)}$ corresponds to customer l in restaurant i. Let $\psi_t^{(i)}$ be the t-th table in restaurant i and let θ_k denote the k-th dish. If n_{itk} is the number of customers in restaurant i seated around the table t and being served dish k, m_{ik} is the number of tables in restaurant i serving the dish k, and K is the number of unique dishes served in the franchise, then

$$\Pr[z_L^{(i)} \in \cdot \,|\, z_1^{(i)}, \ldots, z_{L-1}^{(i)}, Q_0] = \sum_{t=1}^{m_i} \frac{n_{it\cdot}}{L-1+\alpha} \delta_{\psi_t^{(i)}}(\cdot) + \frac{\alpha}{L-1+\alpha} Q_0(\cdot). \tag{16}$$

where $n_{it\cdot} = \sum_k n_{itk}$ and $m_{i\cdot} = \sum_k m_{ik}$. In other words the customer $z_l^{(i)}$ sits at the table indexed by $\psi_t^{(i)}$ with probability proportional to the number of customers $n_{it\cdot}$ already seated there, in which case we set $z_l^{(i)} = \psi_t^{(i)}$, and it sits at a new table with probability proportional to α, in which case we increment $m_{i\cdot}$, set $n_{im_{i\cdot}\cdot} = 1$, draw $\psi_{m_{i\cdot}\cdot}^{(i)}$ from Q_0 and set $z_l^{(i)} = \psi_{m_{i\cdot}\cdot}^{(i)}$. Note that $\psi_{m_{i\cdot}\cdot}^{(i)}$ is drawn from Q_0 and this is the only reference to Q_0 in the predictive (16). In particular, one has

$$\Pr[\psi_t^{(i)} \in \cdot \,|\, \psi_1^{(1)}, \ldots, \psi_{t-1}^{(i)}] = \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \alpha_0} \delta_{\theta_k}(\cdot) + \frac{\alpha_0}{m_{\cdot\cdot} + \alpha_0} G_0(\cdot). \tag{17}$$

where $m_{\cdot k} = \sum_i m_{ik}$ and $m_{\cdot\cdot} = \sum_i \sum_k m_{ik}$. In other words, to table $\psi_t^{(i)}$ it is assigned the dish indexed by θ_k with probability proportional to the number of tables which have previously served that dish in the franchise, in which case we set $\psi_t^{(i)} = \theta_k$, and it is assigned a new dish with probability proportional to α_0, in which case we increment K, draw θ_K from G_0 and set $\psi_t^{(i)} = \theta_K$. Dishes are chosen with probability proportional to the number of tables which have previously served that dish in the franchise. We refer Teh et al. (2006) for additional details.

We have fitted the hierarchical Dirichlet process admixture model (14)-(15) to a set of 188 phased haplotypes from the Colombian (CLM) sample in the 1000 Genomes project. We have considered a collection of 500 bi-allelic loci from chromosomes 2. A value of $K = 3$ covers 99% of the typed loci across individuals. This is in agreement with what is known about Colombian ancestry. Latin America has a well-documented history of extensive mixing between Native Americans and

people arriving from Europe and Africa. This continental admixture, which has occurred for the past 500 years (or about 20-25 generations), gives rise to haplotype blocks. For example, in Figure 2 we show posterior inference for the allocation of loci on a segment of chromosome 2 to one of the three major ancestral populations detected in the sample. The results are based on the Maximum A Posteriori clustering configuration. Notice the mosaic structure of the chromosomes.

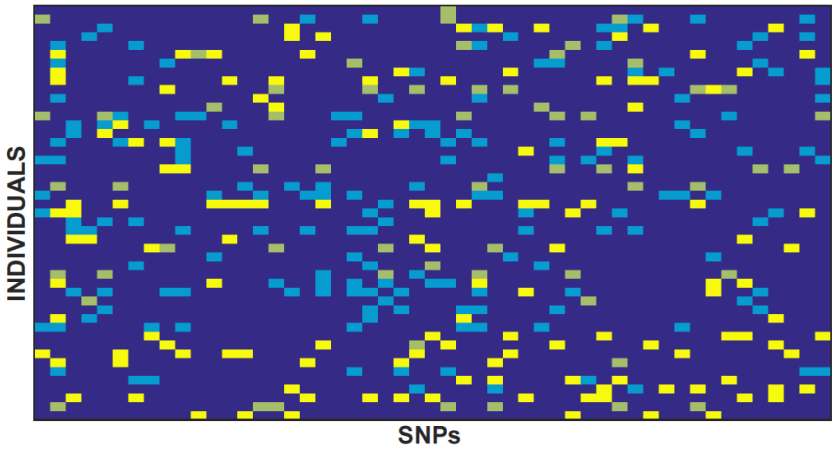

Fig. 2 MAP estimates of population assignment for single nucleotide polymorphism (SNP) data. Each color correspond to a different ancestral population.

Recently, De Iorio et al. (2015) propose the Bayesian nonparametric counterpart of the linkage admixture model defined in (7)-(8), allowing for dependence in the allocation vector $z^{(i)}$. Specifically, let d_l denotes the genetic distance from locus l to locus $l+1$, and let $s_l^{(i)}$ be a binary random variable which denotes whether locus l and locus $l+1$ are on the same segment ($s_l^{(i)}=1$) or not ($s_l^{(i)}=0$). The Bayesian nonparametric linkage admixture model of De Iorio et al. (2015) can be specified as follows

$$
\begin{aligned}
z_1^{(i)} \,|\, Q_i &\sim Q_i \\
s_l^{(i)} &\overset{\text{iid}}{\sim} \text{Bernoulli}(\mathrm{e}^{-d_l r}) \\
z_{l+1}^{(i)} \,|\, z_l^{(i)}, s_l^{(i)}, Q_i &\overset{\text{ind}}{\sim} s_l^{(i)} \delta_{z_l^{(i)}} + (1-s_l^{(i)}) Q_i \\
Q_i \,|\, Q_0 &\overset{\text{iid}}{\sim} \text{DP}(\alpha, Q_0),
\end{aligned}
$$

$$Q_0 \quad \sim \quad \mathrm{DP}(\alpha_0, G_0) \tag{18}$$

and

$$x_l^{(i)} \,|\, Z, P \quad \overset{\text{ind}}{\sim} \quad P_Z$$

$$P_Z \quad \sim \quad \mathrm{Dirichlet}(\lambda_1, \ldots, \lambda_{J_l}), \tag{19}$$

for $i = 1, \ldots, N$ and $l = 1, \ldots, L$. Q_i describes the proportion of the alleles on $x^{(i)}$ coming from each of the populations, as well as the parameters of the populations. Given Q_i, the sequence $x^{(i)}$ is modelled by (i) first placing segment boundaries according to an nonhomogeneous Poisson process with rate $d_l r$ (ii) and then generating alleles on each segment by picking a population according to Q_i and sampling the alleles according to the population specific distribution. The model of Huelsenbeck and Andolfatto (2007) is obtained by letting $r \to +\infty$ and $\alpha \to 0$. By letting $r \to +\infty$ one obtains the standard hierarchical Dirichlet process.

We would like to conclude this section with a note of caution. In particular, within the Bayesian nonparametric settings, inference on K can be sensitive to the choice of the prior on the number of populations, in particular to the prior specification on α and α_0, as well as on the λ_i's. We note that as the number of sequences and/or markers increases the model tends to generate spurious clusters, i.e. clusters with very few individuals in them. This is in agreement with recent results on the clustering properties of the Dirichlet Process. See, e.g., Miller and Harrison (2014) for details. Nevertheless, the number of clusters explaining the majority of the data, i.e. 95-99%, is quite robust to prior specifications. In general, the biological interpretation of K is difficult. See, e.g., Pritchard et al. (2000) and references therein for a detailed discussion. See Fritsch and Ickstadt (2009) for a description of methods for summarizing posterior clustering output.

3.2 The MCMC algorithm

We briefly present the MCMC algorithm for posterior sampling from the Bayesian nonparametric linkage admixture model (18)-(19). The conditional distributions of $(Q_i)_{0 \le i \le N}$, given $(z_l^{(i)})_{1 \le i \le N, 1 \le l \le L}$ and $(s_l^{(i)})_{1 \le i \le N, 1 \le l \le L}$, follow from standard results on the hierarchical Dirichlet process. Conditionally to $(z_l^{(i)})_{1 \le i \le N, 1 \le l \le L}$ and $(s_l^{(i)})_{1 \le i \le N, 1 \le l \le L}$, let K^* be the number of populations. Then,

$$Q_0 = \sum_{j=1}^{K^*} q_{0j} \delta_{\theta_j} + w_0 Q_0'$$

and

$$Q_i = \sum_{j=1}^{K^*} q_{ij}\delta_{\theta_j} + w_i Q_i',$$

for $i = 1,\ldots,N$, where Q_0' is independent of $(q_{01},\ldots,q_{0K^*},w_0)$ and Q_i' is independent of $(q_{i1},\ldots,q_{iK^*},w_i)$. Let n_{ik} be the number of chunks in $z_l^{(l)}$ that are assigned to population k, and let m_{ik} be a random variable such that $m_{ik} = 0$ if $n_{ik} = 0$ and $m_{ik} \in \{1,\ldots,n_{ik}\}$ if $n_{ik} > 0$. Moreover, let us define $n_{0k} = \sum_{1\le i\le N} m_{ik}$. Then, we have

$$(q_{01},\ldots,q_{0K^*},w_0)\,|\,(n_{ik},m_{ik})_{0\le i\le N,1\le k\le K^*} \sim \text{Dirichlet}(n_{01},\ldots,n_{0K^*},\alpha_0), \quad (20)$$

$$Q_0'\,|\,(n_{ik},m_{ik})_{0\le i\le N,1\le k\le K^*} \sim \text{DP}(\alpha_0,G_0), \quad (21)$$

$$\begin{aligned}(q_{01},\ldots,q_{0K^*},w_i)\,|\,(n_{ik},m_{ik})_{0\le i\le N,1\le k\le K^*},(q_{01},\ldots,q_{0K^*},w_0) \quad (22)\\ \sim \text{Dirichlet}(\alpha q_{01}+n_{i1},\ldots,\alpha q_{0K^*}+n_{0K^*},\alpha w_0)\end{aligned}$$

and

$$Q_i'\,|\,(n_{ik},m_{ik})_{0\le i\le N,1\le k\le K^*}, \qquad G_0' \sim \text{DP}(\alpha,G_0'). \quad (23)$$

Equations (20) and (22) form a hierarchy of Dirichlet distributions while Equations (21) and (23) form a hierarchy of Dirichlet processes. The two hierarchies are independent. The reader is referred to Teh et al. (2006) for a detailed account on Equations (20), (21), (22) and (23).

In order to sample from (21) and (23), De Iorio et al. (2015) adopted the slice sampling approach of Walker (2007). See also Papaspiliopoulos and Roberts (2008). The slice sampling allows to truncate the series representations of $Q_0'\,|\,(n_{ik},m_{ik})_{0\le i\le N,1\le k\le K^*}$ and $Q_i'\,|\,(n_{ik},m_{ik})_{0\le i\le N,1\le k\le K^*},G_0'$ while retaining exactness in sampling from them. The idea consists in introducing an auxiliary random variable C_i, the so-called slice variable, such that

$$C_i\,|\,(n_{ik},m_{ik})_{0\le i\le N,1\le k\le K^*}, \quad (Q_i)_{0\le i\le N} \sim \text{Uniform}(0,\min_l q_{iz_l^{(i)}}).$$

Conditionally on C_i, only the atoms with mass above the minimum threshold $\min_i C_i$ need to be simulated. This can be easily achieved by using the stick-breaking representation until the left-over mass falls below the threshold. We refer to De Iorio et al. (2015) for additional details on the implementation of the slice sampling for (21) and (23).

Finally, forward-filtering backward-sampling can be used to update the latent state sequences $(z_l^{(i)})_{1\le i\le N,1\le l\le L}$ and $(s_l^{(i)})_{1\le i\le N,1\le l\le L}$. Note that, conditionally on C_i, only populations with $q_{i,k} > C_i$ will have positive probability of being selected, so that the the forward-filtering backward-sampling is computationally tractable. However, as the random variable C_i depends on the latent state sequences $(z_l^{(i)})_{1\le i\le N,1\le l\le L}$ and $(s_l^{(i)})_{1\le i\le N,1\le l\le L}$, conditioning on C_i introduces complex dependencies among the latent state variables which precludes an exact and efficient forward filtering algorithm. De Iorio et al. (2015) proposed instead to ig-

nore the dependencies caused by the slice variable, and use the resulting efficient forward-filtering backward-sampling as a Metropolis-Hasting proposal. The forward-filtering backward-sampling has a computational scaling of the order $O(LK_i)$, linear in both the number L of loci and potential populations K_i, and it represents the most computationally expensive part of the MCMC algorithm. MATLAB software implementing this MCMC scheme is available at http://BigBayes.github.io/HDPStructure.

4 Discussion and Concluding Remarks

The analysis of population stratification is an increasingly important component of genetic studies. Many different methods have been proposed in the literature and often software implementing such methods has been developed and made publicly available. The main goals of population structure analysis can be summarized as follows: detection of population structure in a sample of chromosomes, estimation of the number of populations present in a sample and consequent assignment of individuals to sub-populations. In the case of genetic admixtures scientific interest focuses on inferring the number of ancestral population to a sample, estimating ancestral population proportions to admixed individuals and identifying the genetic ancestry of chromosomal segments within an individual. No single method is able to deal with the variety of research questions relating to genetic ancestry and it is helpful in applications to use a combinations of approaches.

In this chapter we have reviewed model-based clustering methods for population structure within a Bayesian framework. We have shown how the initial parametric modeling strategies have a natural counterpart in Bayesian nonparametrics, which allows for joint estimation of the number of ancestral populations and the population allocation vector for each individual. In this framework, posterior inference is usually performed through MCMC algorithms. These methods can be used for both haplotype and genotype data, although in the latter case at an extra computational cost. It is in theory straightforward to include further prior information such as geographical locations of sampled chromosomes, ethnicity and phase information. Moreover, it is possible to pre-specify the population of origin of some individuals to aid ancestry estimation for individuals of unknown origin and also to include phenotype information. The inferred clustering structure will generally be sensible and able to explain most of the variability in the data, but clusters will not necessarily correspond to "real" populations and biological interpretation of the number of clusters is often difficult, as pointed out in Pritchard et al. (2000).

Acknowledgements We would like to thank Kaustubh Adhikari for kindly providing the pahsed data and Lloyd Elliott for developing user-friendly MATLAB functions for the linked hierarchical Dirichlet process. Stefano Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406. Yee Whye Teh is supported by the European Research Council (ERC) through the European Unions Seventh Framework Programme (FP7/2007-2013) ERC grant agreement 617411.

References

ALDOUS, D. J. (1985). *Exchangeability and related topics.* Ecole d'ete de probabilites de Saint-Flour, XIII. Lecture notes in Mathematics N. 1117, Springer, Berlin.

ALEXANDER, D.H., NOVEMBRE, J. AND LANGE K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664.

ANDERSON, E.C. AND THOMPSON, E.A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217-1229.

BALDING, D.J. AND NICHOLS, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3-12.

BLACKWELL, D. AND MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353-355.

CORANDER, J., WALDMANN, P. AND SILLANPÄÄ, M.J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367-374.

CORANDER, J., WALDMANN, P., MARTTINEN, P. AND SILLANPÄÄ, M.J. (2004). BAPS2: enhanced possibilities for the analysis of population structure. *Bioinformatics* **20**, 2363-2369.

DAWSON, K.J. AND BELKHIR, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* **78**, 59-77.

DE IORIO, M., ELLIOTT, L., FAVARO, S., ADHIKARI, K. AND TEH, Y.W. (2015). Modeling population structure under hierarchical Dirichlet processes. *Preprint arXiv:1503.08278*.

EVANNO, G., REGNAUT, S. AND GOUDET, J. (2005). Detecting the number of clusters of individuals using the software Structure: a simulation study. *Mol. Ecol.* **14**, 2611-2620.

FALUSH, D., STEPHENS, M. AND PRITCHARD, J.K. (2003). Inference of population structure from multilocus genotype data: linked loci and correlated allele frequencies.. *Genetics* **164**, 1567-1587.

FALUSH, D., STEPHENS, M. AND PRITCHARD, J.K. (2007). Inference of population structure using multi locus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574-578.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209-230.

FIELD, D.L., AYRE, D.J., WHELAN, R.J. AND YOUNG, A.G. (2011). Patterns of hybridization and asymmetrical gene flow in hybrid zones of the rare *Eucalyptus aggregata* and common *E. rubida*. *Heredity* **106**, 841-853.

FRITSCH, A. AND ICKSTADT, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, **4**, 367-392.

GHOSHAL, S. (2010). Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics*, Eds. Hjort, N.L., Holmes, C.C., Müller, P. and Walker, S.G. Cambridge University Press.

HUBISZ, M.J., FALUSH, D., STEPHENS, M. AND PRITCHARD, J.K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resources* **9**, 1322-1332.

HUELSENBECK, J.P. AND ANDOLFATTO, P. (2007). Inference of population structure under a Dirichlet process model. *Genetics* **175**, 1787-1802.

MILLER, J.W. AND HARRISON, M.T. (2014) Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research* **15**, 3333-3370.

NOVEMBRE, J. AND STEPHENS, M. (2008) Interpreting principal components analyses of spatial population genetic variation. *Nature Gentics* **40**, 646–649.

PAPASPILIOPOULOS, O. AND ROBERTS, G.O. (2008). Retrospective Markov Chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169-186.

PARKER, H.G., KIM, L.V., SUTTER, N.B., CARLSON, S., LORENTZEN, T.D., MALEK, T.B., JOHNSON, G.S., DEFRANCE, H.B., OSTRANDER, E.A. AND KRUGLYA, L. (2004). Genetic structure of the purebred domestic dog. *Science* **304**, 1160-1164.

PATTERSON, N., PRICE, A.L. AND REICH, D. (2006) Population structure and eigenanalysis. *PLoS Genetics* **2**, 2074-2093.

PELLA, J. AND MASUDA, M. (2006). The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Can. J. Fish. Aquat. Sci.* **63**, 576-596.

PRITCHARD, J.K., STEPHENS, M. AND DONELLY, P. (2000). Inference on population structure using multilocus genotype data. *Genetics* **155**, 945-959.

NEAL, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* **9**, 249-265.

RANALLA, B. AND MOUNTAIN, J.L. (1997). Detecting immigration by using multilocus genotypes. *Proc. Natl. Acad. Sci.* **94**, 9197-9201.

RAY, A. AND QUADER, S. (2014). Genetic diversity and population structure of *Lantana camara* in India indicates multiple introductions and gene flow. *Plant Biology* **16**, 651-658.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*. **4**, 639-650.

TEH, Y.W., JORDAN, M.I., BEAL, M,J. AND BLEI, D.M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566-1581.

WALKER, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.* **36**, 45-54.

WASSER, S.K., MAILAND, C., BOOTH, R., MUTAYOBA, B., KISAMO, E., CLARK, B. AND STEPHENS, M. (2007). Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences* **104**, 4228-4233.