DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 12: Application of Bayesian Inference in Population Genetics

Lecturer: Clara Wong-Fannjiang

12.1 Recap: Bayesian Inference

In the last few weeks, we have covered several ideas from Bayesian inference.

In Bayesian inference, we have to specify a prior $\mathbb{P}(\theta)$ that reflects our knowledge of or beliefs about the parameter θ before we collect any data. We also have to choose a model $\mathbb{P}(X|\theta)$ that reflects what we believe about the data-generating process. The main goal is to study the posterior, $\mathbb{P}(\theta|X)$, which reflects our updated beliefs about θ after we get to observe the data X.

There are two main approaches for getting $\mathbb{P}(\theta|X)$:

1. Sampling: sampling methods like Markov Chain Monte Carlo (MCMC) are nice because they are asymptotically correct. That is, if you run them, in the limit of infinite time, eventually the samples will be independent samples from the true posterior.
2. Variational inference: this approach, which we will see briefly at the end of this lecture, solves an optimization problem to get at the posterior distribution in lieu of drawing samples.

Today, we will develop the Bayesian model and sampling method behind STRUCTURE, a widely-used population genetics tool. In our derivation of STRUCTURE, we will see some practical weaknesses of sampling, which will motivate a high-level introduction to variational inference methods.

12.2 Application: Inferring Population Structure from Genetic Data

Suppose we have genetic data from N individuals. Can we infer which of K populations these N individuals come from, even without knowing what each of these population looks like beforehand?

It turns out that this problem of inferring population structure from genetic data is incredibly useful for biologists. For instance, population genetics methods are used by researchers to study both modern and historical human migration patterns, as well by companies like 23andMe to provide personal ancestry. They also play an incredibly important role in Genome Wide Association Studies (GWAS), of which we saw one example in Discussion 2. In GWAS, the researcher collects genetic data from thousands of individuals and uses that information to discover genetic marker for particular traits. Population genetics can be used to detect population-specific bias in genetic studies, thereby allowing the researcher to make sure her GWAS results are not biased towards a particular population that's represented in her data.

12.3 Background on Genetics

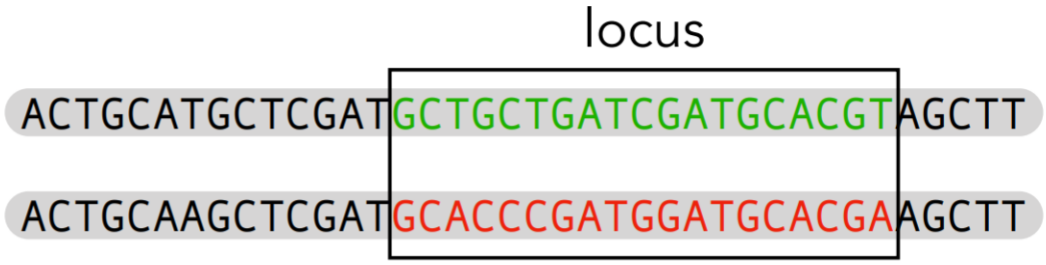

Figure 12.1: Diploid organisms, like humans, have two copies of the genome. A locus is a specific location on the genome.

Our genomes consist of DNA, which is a sequence of nucleotide bases: A, C, T, and G. Humans are diploid organisms, meaning that we have two complete copies of the genome – one inherited from our mothers, and the other from our fathers. A **locus** is a specific location on the genome (e.g. a specific base position, a specific gene, etc.). We have two copies of each locus, as illustrated in Figure 12.1. The term **alleles** is used to refer to all of the possible states that a locus can take on (for example, a single base position has four alleles: A, C, T, and G), and the **genotype** of an individual refers to the two specific alleles that individual has at each of L loci of interest.

12.4 An Initial Hierarchical Model for Population Genetics

Different populations of people tend to have distinctive genotypes. Based on this observation, we would like to try to infer which of K populations N individuals come from (and something about what those populations look like) based on the genotypes of those individuals.

How should we model the different aspects of our problem? As always, whatever model we choose is not going to be a perfect description of reality. In general, there is an important trade-off between model tractability and expressability: a very complicated model that takes into account the influence of many different variables will be very expressive, but also typically much harder to perform inference on (e.g. very slow to sample from using MCMC). One fairly simple mathematical description for our problem is the following:

- We will model a population as a set of L vectors, each of which gives the allele frequencies of a locus. Different populations are thus characterized by having different allele frequencies at the L loci. More concretely, we define p_{kl} to be the vector of allele frequencies at locus l in population k, and therefore p_{klj} gives the frequency of allele j as locus l in population k.
- To model the population of origin of the i^{th} individual, we will use a categorical variable $z^{(i)}$.
- The i^{th} individual's two alleles at loci l will be represented using a pair of variables $(x_l^{(i,1)}, x_l^{(i,2)})$. The individual's genotype consists of L such pairs giving the two alleles at each of L loci.

To ease notation, we define

$$\begin{aligned} P &= \{p_{klj}\}_{k,l,j}, \\ Z &= \{z^{(i)}\}_i, \\ X &= \{(x_l^{(i,1)}, x_l^{(i,2)})\}_{i,l} \end{aligned}$$

where $i = 1, \ldots, N$, $k = 1, \ldots, K$, $l = 1, \ldots, L$, and $j = 1, \ldots, J_l$. Note that each locus l might have a different total number of possible alleles J_l.

We have defined three key quantities: P, the population allele frequencies; Z; the populations of origin of the individuals; and X, the genotypes of the individuals. The relationships between these variables is captured by the graphical model Figure 12.2. We think of P and Z as being independent, as they are just some facts about reality drawn from some prior. However, the observed genotypes X will depend on both P and Z; this captures the intuition that if we know both an individual's population of origin and the allele frequencies for that population, then we have a significant amount of information about what their genome will look like.

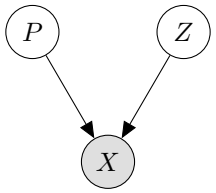

Figure 12.2: Our proposed hierarchical model for population genetics as a graphical model.

In order to actually perform inference, we need to specify distributions for each of these variables. We will assume that, conditioned on the populations of origin Z and the population allele frequencies P, individuals' genotypes are randomly drawn from a multinomial distribution based on the relevant allele frequencies:

$$\begin{aligned} x_l^{(i,1)} &\sim \text{Multinomial}(1, (p_{z^{(i)}l1}, \ldots, p_{z^{(i)}lJ_l})), \\ x_l^{(i,2)} &\sim \text{Multinomial}(1, (p_{z^{(i)}l1}, \ldots, p_{z^{(i)}lJ_l})), \text{ and} \\ \mathbb{P}(x_l^{(i,1)} = j | Z, P) &= p_{z^{(i)}lj}, \text{ for all } j = 1, \ldots, J_l. \end{aligned}$$

Note that this model assumes alleles are distributed independently across different loci. A biologist would argue that this is quite a ridiculous assumption; it is known that genes are different loci are often correlated with one another, for instance due to the fact that we tend to inherit an entire chromosome at a time from our parent. This correlation between loci is called "linkage disequilibrium." We can hope that our simple model is still good enough to be useful, despite the fact that it does not capture linkage disequilibrium.

In Figure 12.2, Z is not being generated by any other variables, so we need only define some prior over Z. We will assume that individuals' populations of origin are drawn independently and

identically according to a multinomial distribution:

$$z^{(i)} \sim \text{Multinomial}\left(1, \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)\right),$$
$$\mathbb{P}(z^{(i)} = k) = \frac{1}{K},\ k = 1, \ldots, K.$$

Again, we can see that this simple model might be inappropriate in certain situations. For example, population sizes are not all the same, so it is unlikely that individuals will be equally likely to come from every population. We have also assumed that individuals are independently distributed, whereas in practice, many members of the same family, clan, etc. take part in the same study. Many statisticians have made careers out of correcting for the effects of family structure in genetics studies; we can nevertheless hope that this model is good enough for us to make useful inferences.

It remains to decide how to model P. Recall that each p_{kl} is a vector of allele frequencies, and thus the entries of p_{kl} must sum to 1. We will assume that the p_{kl} are drawn independently and identically from a Dirichlet distribution. The Dirichlet distribution is a generalization of the Beta distribution to more than two categories, and specifies a distribution over probability vectors (ie. vectors with entries that are all non-negative and sum to 1). Figure 12.3 displays several examples of Dirichlet distributions, and shows how the parameters relate to the location of the mode, skew, and concentration of density.

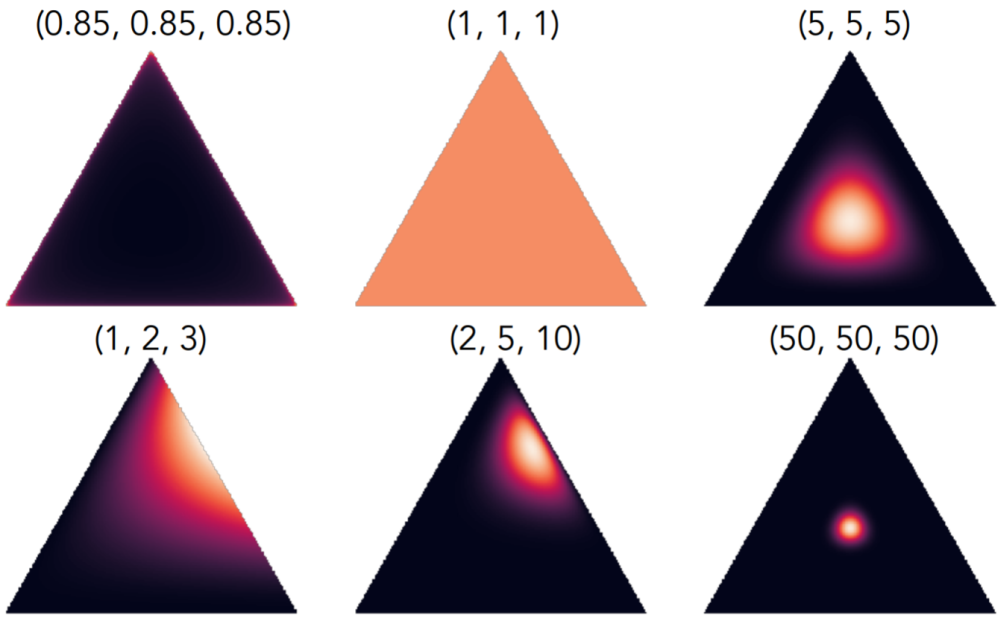

Figure 12.3: Six examples of three-dimensional Dirichlet distributions, each with its parameters given in parentheses at the top. Each triangle represents a three-dimensional probability simplex, with each point in the triangle being a probability vector. Color is used to indicate the amount of probability weight on each probability vector.

We model $p_{kl} \sim \text{Dirichlet}(\lambda_1, \lambda_2, \ldots, \lambda_{J_l})$ with $\lambda_1 = \lambda_2 = \cdots = \lambda_{J_l} = 1$ so that the density is uniform over all probability vectors.

12.5 Inference in our Population Model: Gibb's Sampling

Given data X, how do we get the posterior $\mathbb{P}(P, Z|X)$? We will use Gibb's sampling, which is well-suited to hierarchical models.

As a reminder, Gibb's sampling typically proceeds according to the following steps:

Gibb's Sampler for $\mathbb{P}(\theta|X)$, $\theta \in \mathbb{E}^d$

1. Initialize $\theta^{(0)}$.
2. For $t = 1, 2, \ldots$:
   $\theta_1^{(t)} \sim \mathbb{P}(\theta_1|X, \theta_2 = \theta_2^{(t-1)}, \ldots, \theta_d = \theta_d^{(t-1)})$
   $\theta_2^{(t)} \sim \mathbb{P}(\theta_1|X, \theta_1 = \theta_1^{(t)}, \theta_3 = \theta_3^{(t-1)}, \ldots, \theta_d = \theta_d^{(t-1)})$
   $\ldots$
   $\theta_d^{(t)} \sim \mathbb{P}(\theta_1|X, \theta_1 = \theta_1^{(t)}, \theta_2 = \theta_2^{(t)}, \ldots, \theta_{d-1} = \theta_{d-1}^{(t)})$

The following gives the Gibb's sampler for our model from the previous section:

Gibb's Sampler for $\mathbb{P}(P, Z|X)$

1. Initialize $P^{(0)}, Z^{(0)}$.
2. For $t = 1, 2, \ldots$:
   $p_{11}^{(t)} \sim \mathbb{P}(p_{11}|X, p_{12} = p_{12}^{(t-1)}, \ldots, z^{(1)} = z^{(1)(t-1)}, \ldots)$
   $\ldots$
   $p_{KL}^{(t)} \sim \mathbb{P}(p_{KL}|X, p_{11} = p_{11}^{(t)}, \ldots, z^{(1)} = z^{(1)(t-1)}, \ldots)$
   $z^{(1)(t)} \sim \mathbb{P}(z^{(1)}|X, p_{11} = p_{11}^{(t)}, \ldots, z^{(2)} = z^{(2)(t-1)}, \ldots)$
   $\ldots$
   $z^{(N)(t)} \sim \mathbb{P}(z^{(N)}|X, p_{11} = p_{11}^{(t)}, \ldots, z^{(N-1)} = z^{(N-1)(t)}, \ldots)$

It turns out that we can leverage the (conditional) independence structure in our model to make our sampler more efficient by using Block Gibb's sampling. All the p_{kl} are independent of each other, so we can sample all the P's at the same time rather than sequentially. Likewise, the $z^{(i)}$ are independent of each other, so we can sample the Z's together. Using Block Gibb's sampling, we can specify our algorithm much more simply:

Block Gibb's Sampler for $\mathbb{P}(P, Z|X)$

1. Initialize $P^{(0)}, Z^{(0)}$.
2. For $t = 1, 2, \ldots$:
   $P^{(t)} \sim \mathbb{P}(P|X, Z = Z^{(t-1)})$
   $Z^{(t)} \sim \mathbb{P}(Z|X, P = P^{(t)})$

Our modeling decisions are also nice because Multinomial and Dirichlet are conjugate distributions. Thus, the posterior of P is also a Dirichlet, given by

$$p_{kl}|X, Z \sim \text{Dirichlet}(1 + n_{kl1}, \ldots, 1 + n_{klJ_l}),$$
$$n_{klj} = |\{(i, a) \ : \ x_l^{(i,a)} = j, z^{(i)} = k\}|,$$

and the posterior of Z is also a Multinomial, given by

$$\mathbb{P}(z^{(i)} = k|X, P) = \frac{\mathbb{P}(x^{(i)}|P, z^{(i)} = k)}{\sum_{k'=1}^{K} \mathbb{P}(x^{(i)}|P, z^{(i)} = k')},$$
$$\mathbb{P}(x^{(i)}|P, z^{(i)} = k) = \prod_{l=1}^{L} p_{klx^{(i,1)}} p_{klx^{(i,2)}}.$$

12.6 Admixture and a Revised Hierarchical Model

Is it true that an individual really has just *one* population of origin? If not, our modeling assumptions on $z^{(i)}$ seems overly limiting.

Indeed, it usually does not make sense to think of an individual being from a single population of origin. **Admixture** is when a genotype has multiple populations. We can account for admixture by specifying the proportion of individual i's genotype that originates in population k, denoted $q_k^{(i)}$. Thus, $q^{(i)}$ denotes the vector of population frequencies in the i^{th} individual's genotype, and $Q = \{q_k^{(i)}\}_{k,i}$.

In order to model admixture, we will also need to update the way we represent population of origin. Now, population of origin will be assigned to each allele of each locus of each individual, instead of to each individual. That is, instead of having $z^{(i)}$ for each individual i, we have $(z_l^{(i,1)}, z_l^{(i,2)})$ for each allele at locus l of individual i. Now, $Z = \{(z_l^{(i,1)}, z_l^{(i,2)})\}_{i,l}$.

The graphical model representation of our revised population genetics model is given in Figure 12.4. We also need to define a prior over Q. Since $q^{(i)}$ is a probability vector, it again makes sense to use a Dirichlet distribution. The STRUCTURE authors chose to use $q^{(i)} \sim \text{Dirichlet}(\alpha, \alpha, \ldots, \alpha)$ where α is a hyperparameter that they chose.

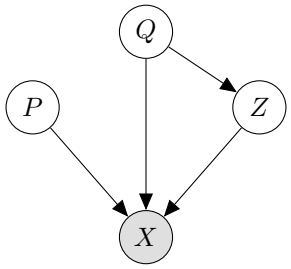

Figure 12.4: Graphical model for the population genetics model which accounts for admixture.

Since all of the Q's are independent of each other, and also of all the P's, we can update our Block Gibb's sampler as follows:

How do we get $\mathbb{P}(P, Q, Z|X)$ Since all the Q's are independent of each other and also independent of all the P's, we can update our Block Gibb's sampler as follows:

Block Gibb's Sampler for $\mathbb{P}(P, Q, Z|X)$

1. Initialize $P^{(0)}, Q^{(0)}, Z^{(0)}$.
2. For $t = 1, 2, \ldots$:
   $P^{(t)}, Q^{(t)} \sim \mathbb{P}(P, Q|X, Z = Z^{(t-1)})$
   $Z^{(t)} \sim \mathbb{P}(Z|X, P = P^{(t)}, Q = Q^{(t)})$

This most recent algorithm is actually exactly the STRUCTURE algorithm. One nice way to interpret the results of using STRUCTURE is to visualize the Z's that we sample. Figure 12.5 gives an example of such a visualization which demonstrates that STRUCTURE picks up the fact that the genotypes of individuals from the same ethic/geographic group are more similar to each other than to those from different groups, even though we do not provide it with any information about where the individuals live or what groups they are a part of.

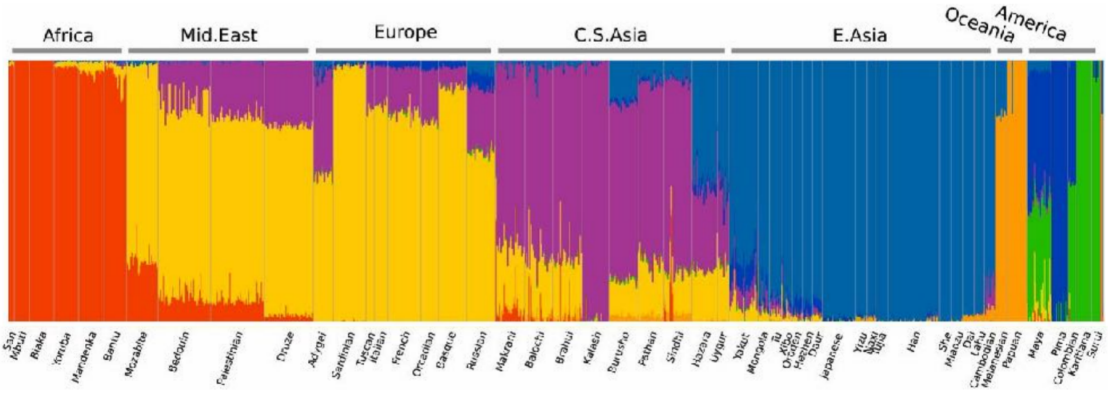

Figure 12.5: Example STRCUTRE output. Each vertical line represents one individual, and the different colors represent different populations of origin. The individuals have been sorted by ethnic/geographic groups, with groups separated by gray lines.

STRUCTURE quickly became the gold standard in genetic and evolutionary biology for inferring population structure from genetic data. However, after a few years, researchers' data sets grew larger and larger (e.g. 100,000's of different loci), and Gibb's sampling quickly became too computationally expensive. People often ended up having to fall back on PCA techiques, which were less principled than STRUCTURE but the only existing techniques fast enough for larger data sets.

What can we do if we want to be Bayesian, but have large amounts of data? Is there anything we can do that is more scalable than sampling? Variational inference, which we discuss briefly in the next section, is one alternative to sampling for performing Bayesian inference.

12.7 An Alternative Approach: Variational Inference

Recall from our discussion of MCMC methods that sampling from the posterior can be difficult, since Markov chains can take a long time to mix in practice, and we cannot always be sure how to assess when they have mixed. The idea behind variational inference is that, instead of sampling, we fit a good approximation of the posterior. While the posterior could be arbitrarily complicated in a general statistical model, we can pick some class Q of simpler distributions that we know how to work with, and find the best approximation of the posterior in Q. Variational inference finds the distribution q^* in Q which is closest to the posterior in terms of some divergence metric, usually the Kullback-Leibler divergence:

$$q^* = \underset{q \in Q}{\operatorname{argmin}} D_{KL}(q(Z)||\mathbb{P}(Z|X)).$$

A few years ago, the authors of the original STRUCTURE method decided to try revamping STRUCTURE to use variational inference instead of Gibb's sampling. They called it fastSTRUCTURE, and, indeed, it is much faster than STRUCTURE while being roughly equally accurate. This population genetics example is thus one illustration of the fact that variational inference can be a very useful alternative approach to Bayesian inference in practice.

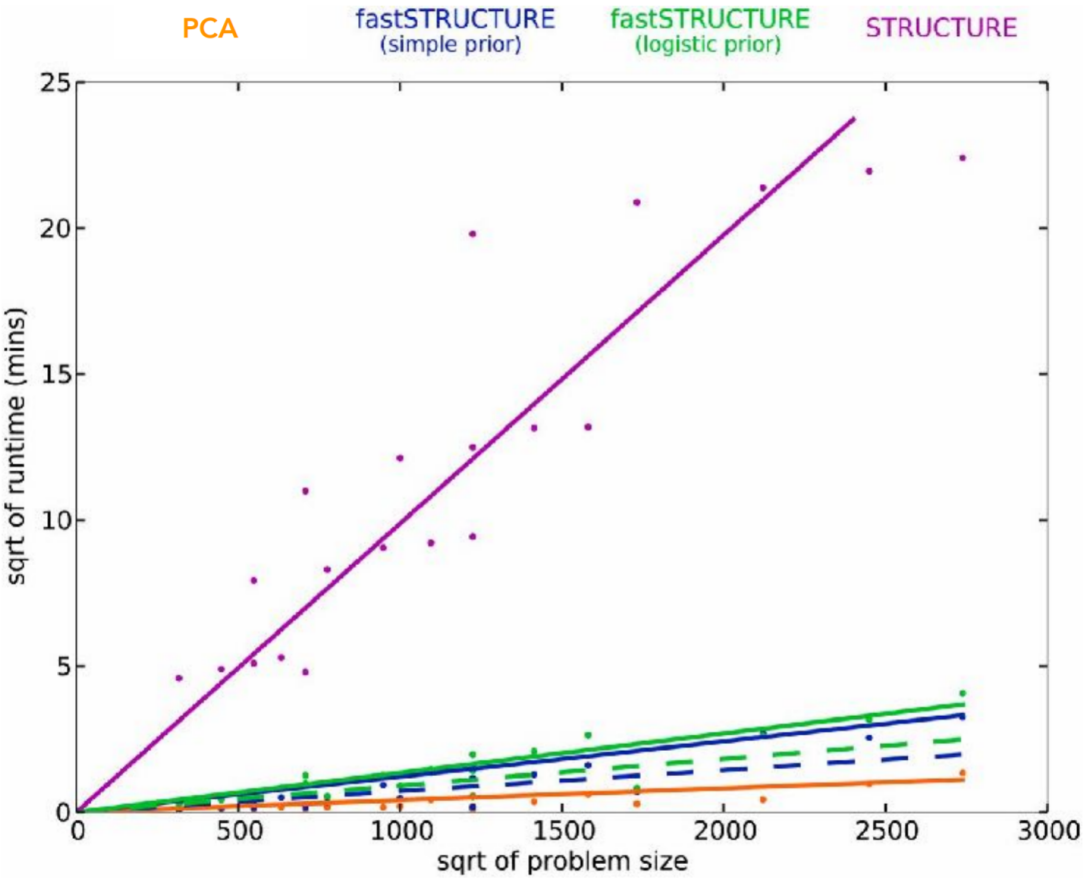

Figure 12.6: A comparison of the runtime of STRUCTURE (sampling-based), fastSTRUCTURE (variational inference), and PCA which also shows how computation time increases with problem size.