LECTURE 14
Latent Dirichlet Allocation

The biological problem we look at is inference of population structure from genetic data. This is an important problem in population genetics/biology both to understand the genetic history of of populations as well to control for population structure when examining genome wide association studies (GWAS). It turns out this same model can be used to model documents, and in the text analysis setting is called topic modeling.

There are several statistical ideas used these ideas include mixture models, Gibb's sampling, and conjugate priors. The general form of a mixture model over multinomial data is called Latent Dirichlet Allocation.

Inference of population structure with no admixture

We first look the case where the observed individuals are not admixed. By this we mean that each individual is drawn from an allele distribution coming from one of $k = 1, ..., K$ ancestral populations. We will then look at the case with admixture where each individual's genome can come from a mixture of of the K ancestral populations.

The quantities that define the problem are

(a) $\{X_1, ..., X_n\}$ – The genotypes of the n individuals. These are n observed variables where for each individual we have $x_\ell^{(i,a)} \equiv (x_\ell^{(i,1)}, x_\ell^{(i,2)})$ = the genotype of the i-th individual at the ℓ-th locus where $i = 1, ..., n$ and $\ell = 1, ..., L$.

(b) $\{Z_1, ..., Z_n\}$ – The population of origin of the i-th individual, z^i = the population form which individual i originated where $z^i = \{1, 2, ...K\}$.

(c) $p_{k\ell j}$ = frequency of allele j at locus ℓ in population k where $j = 1, ..., J_\ell$ is the number of possible alleles at locus ℓ and $k = 1, ..., K$. Note that $p_{z(i)\ell j} = \Pr(x_\ell^{(i,a)} = j \mid Z, P)$

The genotypes X are observed.
The population of origin Z is hidden and must be inferred. The frequency variables P must also be inferred.

From the perspective of conditional probabilities we would like to compute the posterior distribution given a likelihood model for the genotypes (and priors on Z and P)

$$\Pr(Z, P \mid X) \propto \Pr(Z) \times \Pr(P) \times \text{Lik}(X; Z, P), \quad \text{Lik}(X; Z, P) \equiv \Pr(X \mid Z, P).$$

We now build up the problem from the simplest setting to the general setting with no admixture. Pretend that there is only one ancestral population $K = 1$ and that at a locus ℓ we have only two possibile alleles $J_\ell = 2$. The likelihood over n individuals at allele ℓ is

$$\text{Lik}(X_\ell^{1,a}, ..., X_\ell^{(n,a)}; p) \propto \prod_{i=1}^{n} p^{X_\ell^{(i,a)}} (1-p)^{1-X_\ell^{(i,a)}},$$

which is a binomial distribution. The generalization from $J_\ell = 2$ to $J_\ell > 2$ or $X_\ell^{(i,a)} = \{0, 1, 2, ..., J_\ell\}$ corresponds to moving from the binomial distribution to the multinomial distribution

$$\text{Lik}(X_\ell^{(1,a)}, ..., X_\ell^{(n,a)}; \{p_{\ell 1}, ..., p_{\ell J_\ell}\}) \propto \prod_{i=1}^{n} \left[\prod_{j=1}^{J_\ell} p_{\ell j}^{I(X_\ell^{(i,a)}, j)} \right] = \prod_{j=1}^{J_\ell} p_{\ell j}^{S_{\ell j}}$$

where $p_{\ell j}$ is the probability of the j-th allele at locus ℓ with $\sum_{j=1}^{J_\ell} p_{\ell j} = 1$, $p_{\ell j} \geq 0$, $S_{\ell j} = \#\{X_\ell^{(i,a)} = j\}$ is the number of individuals that have allele j at locus ℓ, and $I(X_\ell^{(i,a)}, j) = 1$ if $X_\ell^{(i,a)}$ is the j-th allele and 0 otherwise (this is called the indicator function).

The parameters $P = \{p_{\ell 1}, ..., p_{\ell J_\ell}\}$ are uncertain. These parameters also can be modeled using a probability distribution. We agin first look at the case where $J_\ell = 2$ the binomial case where we have one parameter p. A natural probability distribution to model p is the beta distribution with parameters $\alpha, \beta > 0$ with

$$f(p; \alpha, \beta) \propto p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \leq p \leq 1$$

the case of $\alpha = \beta = 1$ returns the uniform distribution. If we use the beta distribution to set our prior on p and use the binomial likelihood we obtain the following posterior distribution for p give our data

$$\begin{aligned}
\Pr(p \mid X_\ell^{(1,a)}, ..., X_\ell^{(n,a)}) &\propto \text{Lik}(X_\ell^{(1,a)}, ..., X_\ell^{(n,a)}; p) \times f(p; \alpha, \beta) \\
&= \left[\prod_{i=1}^{n} p^{X_\ell^{(i,a)}} (1-p)^{1-X_\ell^{(i,a)}} \right] p^{\alpha-1}(1-p)^{\beta-1} \\
&= \left[p^{S_\ell}(1-p)^{n-S_\ell} \right], \quad S_\ell = \#\{X_\ell^{1,a} = 1\} \\
&= p^{S_\ell+\alpha-1}(1-p)^{n+\beta-S_\ell-1} \\
&= \text{Beta}(S_\ell + \alpha, n - S_\ell + \beta),
\end{aligned}$$

so the posterior distribution is a beta. The beta and binomial are conjugate distributions. In the case of the multinomial the natural distribution on $\{p_{\ell 1}, ..., p_{\ell J_\ell}\}$ is given by a Dirichlet distribution

$$f(\{p_{\ell 1}, ..., p_{\ell J_\ell}\}; \{\alpha_1, ..., \alpha_{J_\ell}\}) \propto \prod_{j=1}^{J_\ell} p_{\ell j}^{\alpha_j - 1}.$$

Using the Dirichlet as the prior we can show that the posterior distribution of the parameters $(\{p_{\ell 1}, ..., p_{\ell J_\ell}\})$ given the genotype is also Dirichlet

$$\begin{aligned}
\Pr(p \mid X_\ell^{(1,a)}, ..., X_\ell^{(n,a)}) &\propto \text{Lik}(X_\ell^{(1,a)}, ..., X_\ell^{(n,a)}; \{p_{\ell 1}, ..., p_{\ell J_\ell}\}) \times f(\{p_{\ell 1}, ..., p_{\ell J_\ell}\}; \{\alpha_1, ..., \alpha_{J_\ell}\}) \\
&= \left[\prod_{j=1}^{J_\ell} p_{\ell j}^{S_{\ell j}}\right]\left[\prod_{j=1}^{J_\ell} p_{\ell j}^{\alpha_j - 1}\right] \\
&= \prod_{j=1}^{J_\ell} p_{\ell j}^{S_{\ell j} + \alpha_j - 1} \\
&= \text{Dir}(S_{\ell 1} + \alpha_1, S_{\ell 2} + \alpha_2, ..., S_{\ell J_\ell} + \alpha_{J_\ell}).
\end{aligned}$$

So at this point we know how to infer the posterior distribution of allele frequencies if we only had one ancestral population, it is given by the Dirichlet distrbution. Obviously this is not so interesting since this case defeats the point of inferring population structure.

We now extend to the case where $K > 2$ where we have real population structure. We introduce a latent variable $Z^{(i)}$ which assigns to each individual a population of origin. This adding a variable is sometimes called augmentation. If we knew $Z^i = k$ we could write out the posterior distribution of the allele frequencies $p_{k\ell j}$ which are the allele frequencies for alleles $j = 1, ...J_\ell$ at locus ℓ for group k with $S_{k\ell j} = \#\{X^{(i,a)} = j, z^{(i)} = k\}$

$$\begin{aligned}
\Pr(p_{k\ell 1}, ..., p_{k\ell J_\ell} \mid Z^{(i)} = k, X^{(1,a,)}, ..., X^{(n,a)}) &\propto \left[\prod_{j=1}^{J_\ell} p_{k\ell j}^{S_{k\ell j}}\right]\left[\prod_{j=1}^{J_\ell} p_j^{\alpha_j - 1}\right], \\
&= \text{Dir}(\alpha_1 + S_{k\ell 1}, ..., \alpha_{J_\ell} + S_{k\ell J_\ell}).
\end{aligned}$$

This gives us a way to sample from the posterior distribution $\Pr(P \mid Z, X)$.

We will show that we can also sample from $\Pr(Z \mid P, X)$. We can write using Bayes' rule

$$\Pr(Z^{(i)} = k \mid X, P) = \frac{\Pr(X^{(i)} \mid P, z^{(i)} = k)}{\sum_{k'} \Pr(X^{(i,a)} \mid P, z^{(i)} = k')},$$

where

$$\Pr(X^{(i,a)} \mid P, Z^{(i)} = k) = \prod_{\ell=1}^{L} p_{k\ell x^{(i,1)}} p_{k\ell x^{(i,2)}}.$$

At this point we know how to draw $\Pr(Z \mid P, X)$ and $\Pr(P \mid Z, X)$. The problem is we want to draw $\Pr(Z, P \mid X)$. There is a way of doing this in many cases using a procedure called Gibb's sampling. The idea behind Gibb's sampling is if I want to sample from a joint distribution $\Pr(Z, X)$ but can only compute the condiitionals $\Pr(Z \mid P)$ and $\Pr(P \mid Z)$ then I can use the following iterative procedure to sample the joint:

(1) Guess a $Z_{(0)}$
(2) For $t = 1$ to T
    (a) sample $P_{(t)} \mid Z_{(t-1)}$
    (b) sample $Z_{(t)} \mid P_{(t)}$
(3) Remove the first t_0 pairs of $(P_{(t)}, Z_{(t)})$, this is called burn-in
(4) Keep every a-th pair of the remaining $(P_{(t)}, Z_{(t)})$, this is called thinning
(5) We now have a iid draws from $\Pr(Z, P)$

The procedure in STRUCTURE adapts the above algorithm in the following way

(1) For $i = 1$ to n: $Z_{(0)}^{(i)} \overset{iid}{\sim} \text{Uni}(1, ..., K)$
(2) For $t = 1$ to T
  (a) For each k, ℓ

$$P_{k\ell\cdot}^{(t)} \mid X, Z_{(t-1)} \sim \text{Dir}(\lambda_1 + n_{k\ell 1}, ..., \lambda_{J_\ell} + n_{k\ell J_\ell}),$$

  where $n_{k\ell j} = \#\{(i, a) : X_\ell^{i,a} = k, Z_{(t-1)}^{(i)} = k\}$
  (b) For each i

$$\Pr\left(Z_{(t)}^{(i)} = k \mid X, P^{(t)}\right) = \frac{\Pr(X^{(i)} \mid P^{(t)}, z^{(i)} = k)}{\sum_{k'} \Pr(X^{(i,a)} \mid P^{(t)}, z^{(i)} = k')},$$

  where

$$\Pr\left(X^{(i,a)} \mid P^{(t)}, Z^{(i)} = k\right) = \prod_{\ell=1}^{L} p_{k\ell x^{(i,1)}} p_{k\ell x^{(i,2)}}.$$

Inference of population structure with admixture

In the case where there is admixture each individual does not necessarily come from one ancestral population. Their genome comes from a mixture of ancestral populations. This is admixture. To model this we need to introduce a new variable Q and adjust our previous variable Z. Our variable P and X are the same as before. The new or adjusted variables are:

(a) $\{Q_1, ..., Q_n\}$ – Vectors of admixture proportions for each individual with

$q_k^{(i)}$ = proportion of i-th individuals genome that originated in population k

(b) $\{Z\}$ – Allele copy $X_\ell^{(i,a)}$ originated in unknown population $Z_\ell^{(i,a)}$

$$z_\ell^{(i,a)} = \text{population of origin of allele copy } X_\ell^{(i,a)}$$

note previously we only needed one Z for each individual.

We observe that

$$\Pr(X_\ell^{(i,a)} = j \mid Z, P, Q) = p_{z_\ell^{(i,a)} \ell j},$$

and

$$\Pr(z^{(i,a)} = k \mid P, Q) = q_k^{(i)},$$

and we can place the prior

$$q^{(i)} \sim \text{Dir}(\alpha, ..., \alpha).$$

We will see soon we can write the following conditionals

$$P, Q \mid X, Z, \quad Z \mid X, P, Q.$$

This lets us write out the following Gibbs sampler.

(1) For each i, a: $Z_{(0)}^{(i,a)} \overset{iid}{\sim} \text{Uni}(1, ..., K)$
(2) For $t = 1$ to T
  (a) For each k, ℓ

$$P_{k\ell\cdot}^{(t)} \mid X, Z_{(t-1)} \sim \text{Dir}(\lambda_1 + n_{k\ell 1}, ..., \lambda_{J_\ell} + n_{k\ell J_\ell}),$$

  where $n_{k\ell j} = \#\{(i, a) : X_\ell^{i,a} = k, Z_{(t-1)}^{(i,a)} = k\}$

(b) For each i

$$q_{(t)}^{(i)} \mid X, Z_{(t-1)}^{(i,a)} \sim \mathrm{Dir}(\alpha + m_1^{(i)}, ..., \alpha + m_k^{(i)},$$

where

$$m_k^{(i)} = \#\{(\ell, a) : z_\ell^{(i,a)} = k\}.$$

(c) For each i, a, ℓ

$$\Pr\left(Z_{(t)}^{(i,a)} = k \mid X, P^{(t),Q_{(t)}}\right) = \frac{q_k^{(i)} \Pr(X_\ell^{(i)} \mid P^{(t)}, z^{(i)} = k)}{\sum_{k'} q_k'^{(i)} \Pr(X_\ell^{(i,a)} \mid P^{(t)}, z^{(i)} = k')},$$

where

$$\Pr\left(X_\ell^{(i,a)} \mid P^{(t)}, Z^{(i)} = k\right) = p_{k\ell x^{(i,a)}}.$$