Project 2: Latent Dirichlet Allocation

Kennard Ng and Luo Wen Han

November 9, 2019

1 Introduction

In this report, we make inferences on the ancestry of individuals given their genotype. More specifically, we model the relationship between a person's genotype and his ancestors using the Latent Dirchlet Allocation (LDA) model [1] and use variational inference to discover the ancestor populations' of individuals.

2 Background

2.1 Notations and Terminology

In our problem, we define the following:

- a genotype $w \in \{w^v\}_{1:V}$ is an atomic unit in genetics where each genotype is represented as a one-hot encoded vector of length V where V is the number of genotypes.
- an individual is represented as a sequence of N genotypes $\mathbf{w} = \{w_n\}_{1:N}$.

2.2 Latent Dirichlet Distribution

Intuitively, each individual's sequence of genotypes are inherited from his ancestors where under the generative probabilitistic model of LDA model, individuals can be represented as random mixtures over latent ancestor populations, where each ancestor population can be described by a distribution over genotypes. More specifically, the generative process for each individual is the following:

1. Choose $\theta \sim \text{Dir}(\alpha)$.
2. For each of the N genotypes w_n:
   (a) Choose an ancestor population $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a genotype w_n from $p(w_n|z_n, \beta)$, where $p(w_n|z_n, \beta)$ is a multinomial distribution conditioned on the ancestor population z_n and β,

where β is $k \times V$ matrix of genotype probabilities conditioned on topic $z \in \{z_k\}_{1:K}$ i.e. $\beta_{ij} = p(w^i|z_j)$ and α k-vector parameter with components $\alpha_i > 0$ such that:

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma\left(\alpha_i\right)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \tag{1}$$

where $\gamma(\cdot)$ is the Gamma function. The graphical model representation of the LDA model is given in Figure 1.

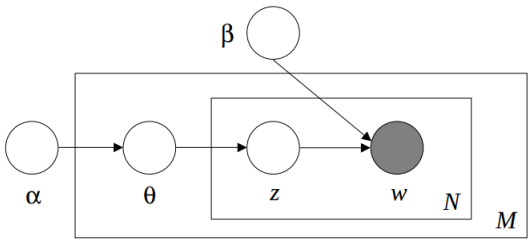

Figure 1: Graphical Model of LDA. Taken from [1]. The plates represent replicas. In our context, the inner plate represents the repeated choice of ancestor populations and genotypes of a person while the outer plate represents the population with M individuals. Notice also that by observing w, β and θ are coupled together (head-to-head relationship).

2.3 Variational Inference

During inference, we want to find the posterior distributions over hidden variables $\theta, \mathbf{z}$ given parameters α, β and observed variables $\mathbf{w}$. We express this as the following posterior distribution:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}. \tag{2}$$

Unfortunately, we see that exact inference is intractable given that the normalizer is intractable where marginalizing over the hidden variables gives:

$$\begin{aligned} p(\mathbf{w}|\alpha, \beta) &= \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p\left(z_n|\theta\right) p\left(w_n|z_n, \beta\right) \right) d\theta \\ &= \frac{\Gamma\left(\Sigma_i \alpha_i\right)}{\prod_i \Gamma\left(\alpha_i\right)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} \left(\theta_i \beta_{ij}\right)^{w_n^j} \right) d\theta, \end{aligned} \tag{3}$$

which is intractable due to the coupling between β and θ. Hence, we instead use approximate inference where, under the Mean Field assumption, we remove the problematic edges between $\theta, \mathbf{z}$ and $\mathbf{w}$, the node $\mathbf{w}$ and endow our model with free variational parameters, introducing the following variational distribution:

$$q(\theta, z \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n). \tag{4}$$

where the Dirchlet parameter γ and the multinomial parameters $\{\phi_n\}_{1:N}$ are free variational parameters. The graphical model representation of the variationl distribution based on the mean field assumption is given in Figure 2.

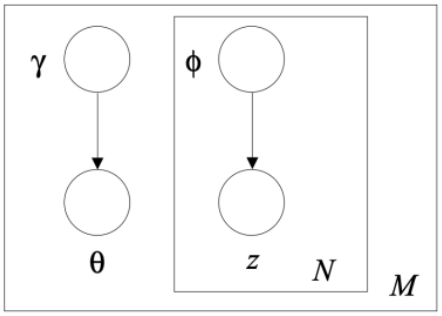

Figure 2: Approximate Distribution under the Mean Field Assumption.

Blei et al. [1] show that the optimal values of the variational parameters γ and ϕ can be computed by minimizing the Kullback-Leibler (KL) divergence between the approximate distribution and the true posterior where:

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \mathbf{KL}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)) \tag{5}$$

We minimize the KL-divergence via an iterative fixed-point method using the following update equations dervied by taking the derivatives of the KL divergence w.r.t. the variational parameters and setting them to zero:

$$\phi_{ni} \propto \beta_{iw_n} \exp\left\{\mathbb{E}_q\left[\log\left(\theta_i\right) \mid \gamma\right]\right\}, \tag{6}$$

$$\gamma_i = \alpha_i + \textstyle\sum_{n=1}^{N} \phi_{ni}, \tag{7}$$

where the expectation in the multinomial update in 6 can be computed by:

$$\mathbb{E}_q\left[\log\left(\theta_i\right) \mid \gamma\right] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) \tag{8}$$

where Ψ is the first derivative of the $\log\Gamma$ function computed from Taylor approximation. The iterative fixed point method is summarized to be following:

1. $\forall i, n, \quad \phi_{ni}^0 = 1/k$ and $\forall i, \quad \gamma_i = \alpha_i + N/k$.
2. repeat until convergence:
   (a) for $n \in \{1, \cdots, N\}$:
      i. for $i \in \{1, \cdots, k\}$: $\phi_{ni}^{t+1} = \beta_{iw_n} \exp\left(\Psi(\gamma_i^t)\right)$
      ii. normalize ϕ_{ni}^{t+1} to sum to 1.
   (b) $\gamma^{t+1} = \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$

3 Experiments

In our experiments, we set the number of ancestor populations $K = 4$ and use previously inferred β matrix provided. Our dataset consist of $M = 100$ indivduals represented by a vocabulary of $V = 200$ genotype loci.

We perform variational inference on the dataset to determine the population mixture θ and the genotype ancestry assignments z for individuals in our dataset.

In our first experiment, we perform variational inference on individual 1 and store the values of ϕ and store it in `phi1.out`. We also perform variational inference on all individuals and store the values of θ in `Theta.out`.

Finally, we study how varying α controls the rate of convergence, where convergence happens when both γ, ϕ have an absolute change smaller than $1e^{-3}$. We perform experiments over $\alpha = \{0.01, 0.1, 1.0, 10.0\}$ for each individual and plot the distribution over the number of iterations to convergence α in Figure 3. We also plot the perform inference for each value of α for 100 iterations and present the average time taken for convergence in Table 1.

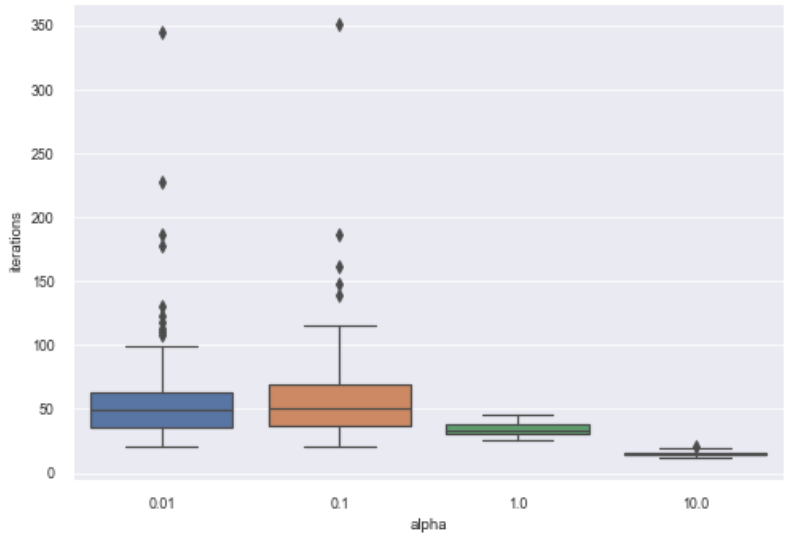

Figure 3: Box plot of number of iterations for each α value

alpha	0.01	0.1	1.0	10.0
time (s)	23.40	22.52	13.47	**6.09**

Table 1: Average time to convergence over 100 iterations

Discussions. We see that larger values of α results in a smaller time to convergence. This is also seen in Figure 3 where the distribution of iterations has a smaller, interquartile range and median for larger values of α.

To that end, we observe the Dirchlet distribution over different values of α for $K = 3$ in Figure 4. We observe that the values of θ are more concentrated given larger α. This implies that θ and from Equation 8, ϕ_{ni} does not vary much across iterations with larger α. As such, convergence happens quickly given that ϕ does not change much across iterations and so does γ.

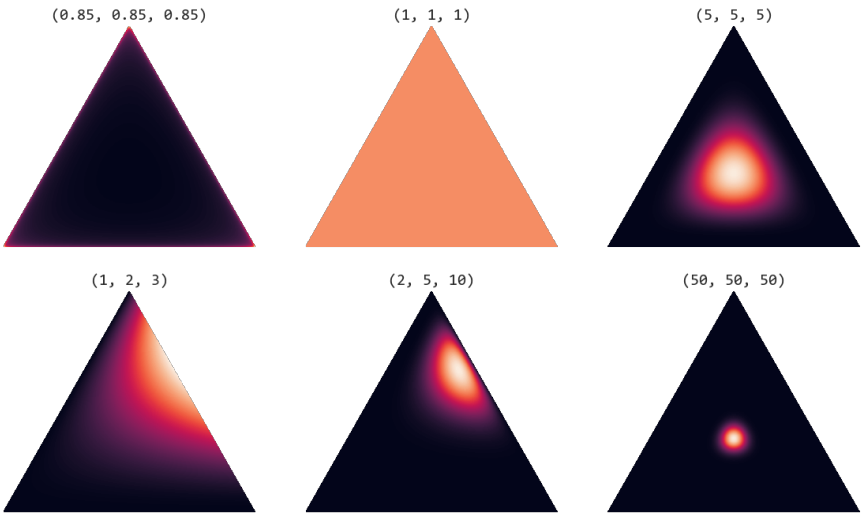

Figure 4: Dirchlet Distribution with different values of α.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov 1999.