

Problem 1.

In this problem, we will apply LDA to human ancestry discovery. In applications of population genetics, it is often useful to classify individuals in a sample into populations. An underlying assumption is that there are K ancestor populations, and each individual is an admixture of the ancestor populations. For each individual, we measure some genetic data about them, called genotype data. Each genotype is a locus that can take a discrete count value, individuals with similar genotypes are expected to belong to the same ancestor populations. We can derive the admixture coefficients θ for each individual by running an LDA model, where the documents are individuals, and the words are the genotype.

Now let us assume the β matrix is known, and focus on variational inference of the population mixture θ and the genotype ancestry (topic) assignments z for any individual. The variational distribution used to approximate the posterior (for each individual) is

$$q_i(\theta, z|\gamma, \phi) = q(\theta_i|\gamma_i) \prod_{n=1}^{N_i} q(z_{in}|\phi_{in}), \quad i = 1, \dots, M$$

The data matrix provides data about $M = 100$ individuals, each represented by a vocabulary of $N = 200$ genotype loci. This data has been preprocessed into a count matrix D of size $M \times N$. D_{ij} is the number of occurrences of genotype j in individual i , and $\sum_j D_{ij}$ is the number of genotype loci in an individual. We learnt the LDA topic model over $K = 4$ ancestor populations, and the data matrix and the known β matrix can be downloaded from the course website. The value of α is 0.1. You may use the following code to load the data in python.

```
1 import pickle
2
3 with open("lda_data.p", "rb") as handle:
4     data_loaded = pickle.load(handle)
```

- (1) Derive the variational inference update equations for estimating γ and ϕ .
- (2) For individual one, run LDA inference to find ϕ for each genotype locus, store it as a matrix of size $n_1 \times K$ (where $n_1 : \sum_{1j} I(D_{1j} \neq 0)$, $I(\cdot)$ being the indicator function, is the number of non-zero genotypes present in individual 1), and plot it as an image in your write up. Don't forget to show the colormap using the `colorbar` function to allow the colors in the image to be mapped to numbers!
- (3) We will construct a matrix Θ of size $M \times K$ to represent the ancestor assignments for all individuals in the population. For each individual i , run LDA inference to find γ , and store it as row of Θ , i.e. $\Theta_i = \gamma$. Visualize Θ as an image.

(4) Report the number of iterations needed to get to convergence for running inference on all M individuals (you may use absolute change less than $1e-3$ as the convergence criteria).

(5) Repeat the experiment for $\alpha = 0.01, \alpha = 1, \alpha = 10$, and for each of α , visualize the Θ matrix summarizing the ancestor population assignments for all individuals. Discuss the changes in the ancestor population assignments to the individuals as α changes. Does the mean number of iterations required for convergence for inference change as α changes?

Problem 2.

A simple first-order autoregressive process, AR(1), is defined as follows:

$$y_t = a + by_{t-1} + \epsilon_t, \quad t \geq 1, \quad y_0 = a,$$

where a and b are some constants and $\epsilon_t \sim \mathcal{N}(0, 1)$ is a Gaussian noise. AR(1) defines a distribution over sequence of discrete values, $\{y_0, y_1, \dots\}$ (to sample from this distribution, you can simply run the forward autoregressive recursion).

Derive a mean, $\mu(t)$, and a kernel, $k(t, t')$, functions for a Gaussian process that defines a distribution over functions, $y(t)$, that coincides with AR(1) for all $t \geq 1$.

Problem 3.

Let G_0 be a distribution over Θ and let α be a positive scalar. For any finite, measurable partition A_1, \dots, A_r of Θ , G is defined to be a Dirichlet process with base distribution G_0 and concentration parameter α_0 , denoted by $G \sim \text{DP}(\alpha_0 G_0)$, if

$$(G(A_1), \dots, G(A_r)) \sim \text{Dir}(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_r)).$$

Suppose we have observation X_1, \dots, X_n , which we assume are drawn from G . Assuming we have the prior $G \sim \text{DP}(\alpha_0 G_0)$, derive the posterior distribution for $G | X_1, \dots, X_n$.

Problem 4.

Consider the following DP mixture of Gaussian model in \mathbb{R}^2

$$\begin{aligned} y_i | \theta_i &\sim \mathcal{N}(\theta_i, \sigma_y^2 I_2) \\ \theta_i | G &\sim G, \quad i = 1, \dots, n \\ G &\sim \text{DP}(\alpha_0 G_0) \\ G_0 &= \mathcal{N}(0, \sigma_0^2 I_2) \end{aligned}$$

Let $\sigma_0 = 5, \sigma_y = 1$. The data on the course website were generated by sampling from this model for a particular choice of α_0 .

(1) Generate 1000 samples from the model with $\alpha_0 = 0.1, 1, 5, 10$ respectively, and show

the scatter plots of your samples.

(2) Download the data from the course website. Implement a collapsed Gibbs sampler for this model in which the θ parameter have been integrated out. Fix $\alpha_0 = 1$. Run the sampler, show the scatter plots of the data and the samples of the unique ϕ_i 's at the first few iterations (e.g., 1, 5, 10, 20, 50). When the sampler appears to have converged, use the subsequent samples to plot a histogram of the posterior distribution of the number of occupied tables.

(3) Now place a vague gamma prior on α_0 . Again plot a histogram of the posterior distribution of the number of occupied tables. Also plot a histogram of the posterior distribution of α_0 . Explore the sensitivity of your results to the choice of parameters for the gamma distribution.

(4) Interpret your results.