Statistical Models and Computing Methods, Problem Set 3

November 17, 2024 Due 12/02/2024

Problem 1.

A total of n instruments are used to observe the same astronomical source. Suppose the number of photons recorded by instrument j can be modeled as $y_j \sim \text{Poisson}(x_j\theta + r_j)$ where $\theta \geq 0$ is the parameter of interest, and x_j and r_j are known positive constants. You may think of θ, x_j, r_j as the source intensity, the observation time, and the background intensity for instrument j, respectively. Assume the photon counts across different instruments are independent.

(1) Write down the likelihood function for θ. *(5 points)*

(2) Introduce mutually independent latent variables $z_{j1} \sim \text{Poisson}(x_j\theta)$ and $z_{j2} \sim \text{Poisson}(r_j)$ and suppose we observe only $y_j \equiv z_{j1} + z_{j2}$. Under this formulation, derive an EM algorithm to find the MLE of θ. *(10 points)*

Table 1: Data (x_j, r_j, y_j)

x	1.41	1.84	1.64	0.85	1.32	1.97	1.70	1.02	1.84	0.92
r	0.94	0.70	0.16	0.38	0.40	0.57	0.24	0.27	0.60	0.81
y	13	17	6	3	7	13	8	7	5	8

(3) Apply your EM algorithm to the data set given by Table 1. What is the MLE? *(10 points)*

(4) For these data compute the observed Fisher information and the fraction of missing information. (Recall the observed Fisher information is defined as the negative second derivative of the observed data log-likelihood evaluated at the MLE.) *(5 points)*

Problem 2.

Let $x_1, \ldots, x_m$ be i.i.d. sample from a normal distribution with mean μ and variance σ^2. Suppose for each x_i we observe $y_i = |x_i|$ rather than x_i. Download the data from the course website.

(1) Derive an EM algorithm to find the MLE of μ and σ^2. *(10 points)*

(2) Apply your EM algorithm to the data with different starting values. Does your EM always converge to the same point estimate? If not, do you observe any pattern of your estimates? Explain it. *(5 points)*

(3) Derive the gradient of the parameters. Compare the standard gradient descent method to EM. Show $\ell^* - \ell$ as a function of the number of iterations (ℓ is the log-likelihood function and ℓ^* is the optimal value of it) for both methods. Which one is better in this case? Why? *(10 points)*

Problem 3.
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

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

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.

```
import pickle

with open("lda_data.p", "rb") as handle:
    data_loaded = pickle.load(handle)
```

(1) Derive the variational inference update equations for estimating γ and ϕ. *(10 points)*

(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! *(10 points)*

(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. *(10 points)*

(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 points)*

(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? *(10 points)*