Homework 2
Yanran Li (yl5465)
November 2, 2023

1 Problem 1

1.1 Introduction

A common design for estimating the concentrations of compounds in biological samples is the serial dilution assay, in which measurements are taken at several different dilutions of a sample, giving several opportunities for an accurate measurement[1].

1.2 Data

- **Description of the dataset:** The data I used to develop the mixture model through Gibbs sampling was in the context of literal contamination that arises in laboratory measurement. Epidemiologic investigations of environmental-disease associations rely on accurate measurements of environmental exposures. To determine the level of environmental exposure, samples are collected, and concentrations are measured in labs using immunoassays with serial dilutions. An immunoassay is a biochemical test that measures the presence or concentration of an analyte depending on the reaction to an antibody or an antigen[2]. Developed in the 1970's, the Enzyme-Linked Immunosorbent Assay (ELISA) is a staple of biological laboratory techniques used to detect and quantify proteins, and other substances that can be bound to antibodies. Typically, the test is conducted with serial dilutions of both a standard sample and multiple unknown samples in a microtiter plate to detect and quantify proteins and other substances that can be bound to antibodies or antigens. Figure 1 shows the two standard dilution curves.

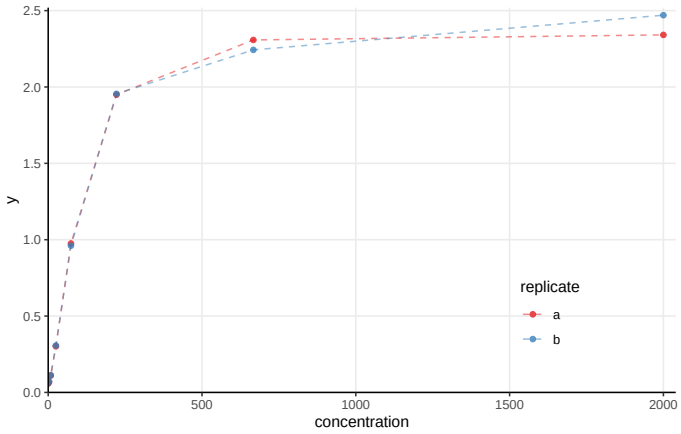

Figure 1: Serial dilution standard curve

- **Description of the features:** Figure 1 shows standards data for the allergen from dust mites, Der f 1 ($\mu g/ml$) in a microtiter plate. The standard data are presented with signal responses (y) and true concentrations ("concentration" of Der f 1 ($\mu g/ml$) as x-axis).

- **Description of the response variable:** In the standard serial dilution process, multiple dilutions are applied to both the standard and unknown samples. Signal responses (y) to the antibody-antigen interaction are measured at the dilutions of each sample. The serial dilutions of the standard sample are used to make a calibration curve relating the signal response to the known concentrations of the diluted standard sample. Here we used the standard signal responses read from the machines as our response variable.

- **Number of observations:** We consider 4 sample plates here and each microtiter plate contains 96 wells (12 columns and 8 rows). Each of the standard sample was prepared at the known and fixed concentration and then diluted. Each of the unknown samples was analyzed using 3 dilutions at 1/10, 1/100, and 1/10000. In this homework situation, we only consider the standard samples (384 observations in total) to implement Gibbs sampler. The samples' signal responses distribution are shown in figure 2

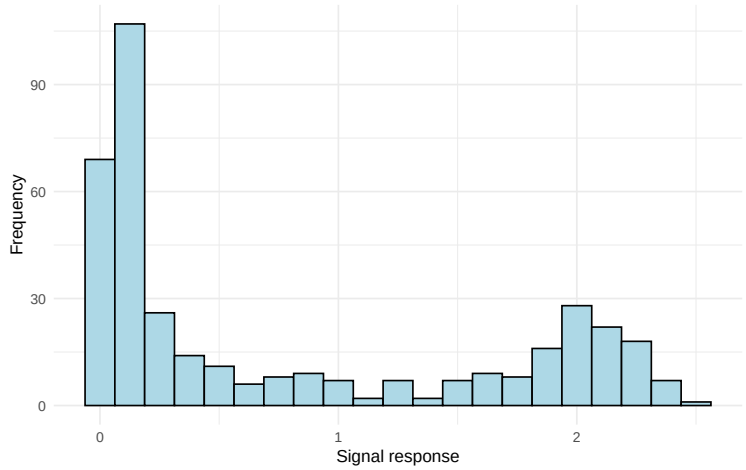

Figure 2: Histogram of signal response

1.3 Model

1.3.1 Gibbs Sampler

Seeing from figure 2, it appears that these data exist in 2 separate clusters and it can fit the Gaussian Mixture assumption. We want to develop a method for finding these latent clusters.

I implemented the Gibbs Sampling on this dataset and recorded the log joint to check the convergence. Specifically, the prior for each cluster' μ implicitly assumed to be a normal distribution, centered around the mean of the data and with the standard deviation of the data. The prior for σ is implicitly assumed to be the same for all clusters and is determined by the standard deviation

of the data. The prior for the latent variable assignments (z) is uniformly distributed over the 2 clusters. Then, my Gibbs sampler iteratively samples from the conditional distribution of each parameter given the others, ultimately providing an approximation of the posterior distribution of these parameters given the data. Through the procedures, I also recorded the log $p(z_{1:n}, \mu_{1:K}, \sigma_{1:K})$ (log joint probability) as a function of iteration and plotted them on figure 3. Seeing from the plot, the Gibbs sampler converges pretty well.

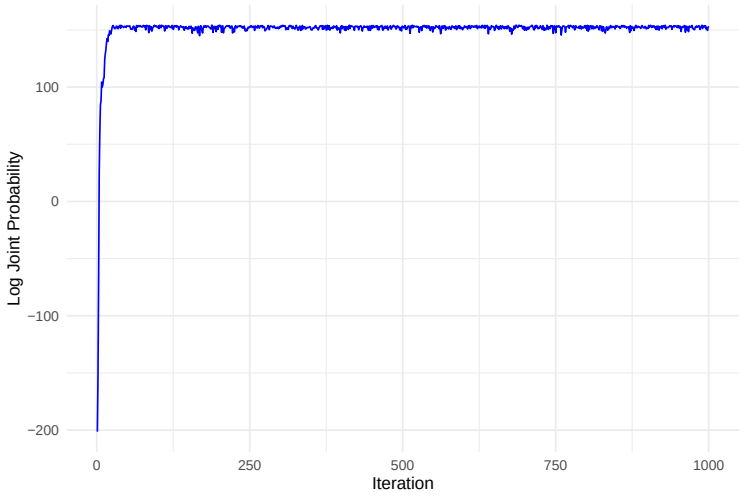

Figure 3: Log joint probability over iterations

1.3.2 Variational Inference

I also implemented variational inference to fit a Gaussian Mixture Model (GMM) on my dataset: for each data point i and cluster k, I compute a term that represents the lower bound on the log likelihood of the data. This term is calculated using the responsibilities ϕ, the cluster parameters (μ and σ^2), and the mixing coefficient (assumed to be 0.5). And then I summed up these terms across all data points and clusters to compute the ELBO for each iteration. I also plotted the ELBO as a function of iteration in figure 4, depicting the convergence of the inference procedure.

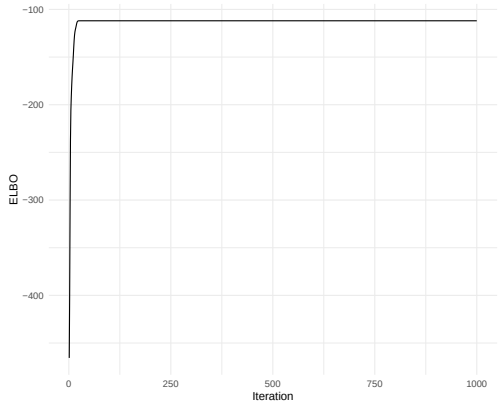

Figure 4: Log joint probability over iterations

2 Problem 2 (Abstract)

Many applied statistical applications face the potential problem of model contamination and measurement error. In this paper, we focus on two challenges: contamination can happen at multiple levels of an experiment, and the model for the contaminated observations is typically itself speculative. We will address these issues by using a multilevel model with application to serial dilution assay, a problem where the current approach can lead to noisy estimates and difficulty in estimating very low or high concentrations or identifying problematic observations. We will propose an Bayesian framework to simultaneously flag problematic observations and estimate concentrations of unknown samples in serial dilution assay with a measure of estimation uncertainty. Our approach will be validated through an analysis of real-world immunoassay data, sourced from the New York City Neighborhood Asthma and Allergy Study, as well as through simulation studies.This research holds significance for researchers to precisely quantify both the nature and extent of contamination and measurement errors.

Key Words: Bayesian inference; Uncertainty quantification; Immunoassay; Measurement error.

References

[1] Andrew Gelman, Ginger Chew, and Michael Shnaidman. Bayesian analysis of serial dilution assays. *Biometrics*, 60:407–17, 07 2004. doi: 10.1111/j.0006-341X.2004.00185.x.

[2] Sandeep Kumar Vashist. *Handbook of Immunoassay Technologies: Approaches, Performances, and Applications*. 01 2018. ISBN 9780128117620.