STATS 451 Final Project Proposal: Method

Yuzhou Peng

April 14, 2024

1 Introduction

Given the dataset we have described in previous section, we have observed birth rates of each counties across all states of the USA. We are interested in modeling the underlying distribution of birth rates in each states.

2 Notation

We define our notations to be used as follow:

- y_{ij} is the observed birth rates of county i in state j.
- ω_j, κ_j are the parameters of underlying distribution of birth rate in state j, which are our quantities of interest.
- ω, κ are the hyper parameters that follow a pre-designed prior distribution

3 Model Construction

To preserve the potential dependency structure within the dataset, we apply Bayesian hierarchical model to estimate the distribution of birth rates in each states. The basic model structure is as follow:

$$y_{ij} \sim Beta(\omega_j(\kappa_j - 2) + 1, (1 - \omega_j)(\kappa_j - 2) + 1);$$
$$\omega_j \sim Beta(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1) \text{ and } \kappa_j \sim Gamma(S, R)$$
$$\omega \sim Beta(A, B) \text{ and } \kappa \sim Gamma(S, R)$$

Where A,B,S and R are pre-designed parameters.

Specifically, to study the two proposed problems, we design two slightly different set of models respectively.

3.1 Inductive Bayesian hierarchical model

To model the data from 2011 to 2016, instead of build the model separately, we want to use the inductive property of Bayesian model. In the 1st year, we assume the prior distribution of hyper parameters to be non-informative and generic vague. After that, The prior distribution of hyper parameters in i-th year is designed to be the estimated marginal posterior distribution of hyper parameters obtained in (i-1)-th year.

3.2 Comparison between full model and partially independent model

In this section, we geographically classified states in USA into several categories, say, West, East Coastal, South, etc. We want to study if there is any difference when we separately apply Bayesian hierarchical model to different partition of the states in USA comparing with fitting all states at one time.

4 Potential Difficulty

1. The estimation of marginal posterior of hyper parameters is difficult since we cannot have an analytic form of this distribution.