Multi-level_LCM

Tian

6/22/2021

1. Introduction

In this simulation, we hope to demonstrate the perfomance of multilevel latent class model on a clustered dataset.

2. Simulation setting

In the dataset, the N observations belong to G clusters; for each observation, they have J causes $A_1, ...A_J$, and they belong to a latent class U which has K categories. We assume the latent class U, given cluster G, follows a categorical distribution:

$$U_i \mid G_i = g \sim \text{ Mutinomial } (\vec{\pi}^g)$$

The $\vec{\pi}^g$ follows a dirichlet prior distribution

$$\vec{\pi}^g \sim \text{ Dirichlet } (\alpha_1, \cdots \alpha_K)$$

$$\text{seting } \alpha_1 = \cdots = \alpha_K = 1$$

As for the causes, given the latent class U, they follow a bernoulli distribution (in more complex setting, another categorical distribution)

$$A_{ij} \mid U_i = k \sim \text{ Bernoulli } \left(p_j^k\right)$$

We aim to estimate $\Pi_{G\times K}$ and $P_{J\times K}$. The violin plot below showed the density plot of samplers and the bar showed the 95% credible interval with median.

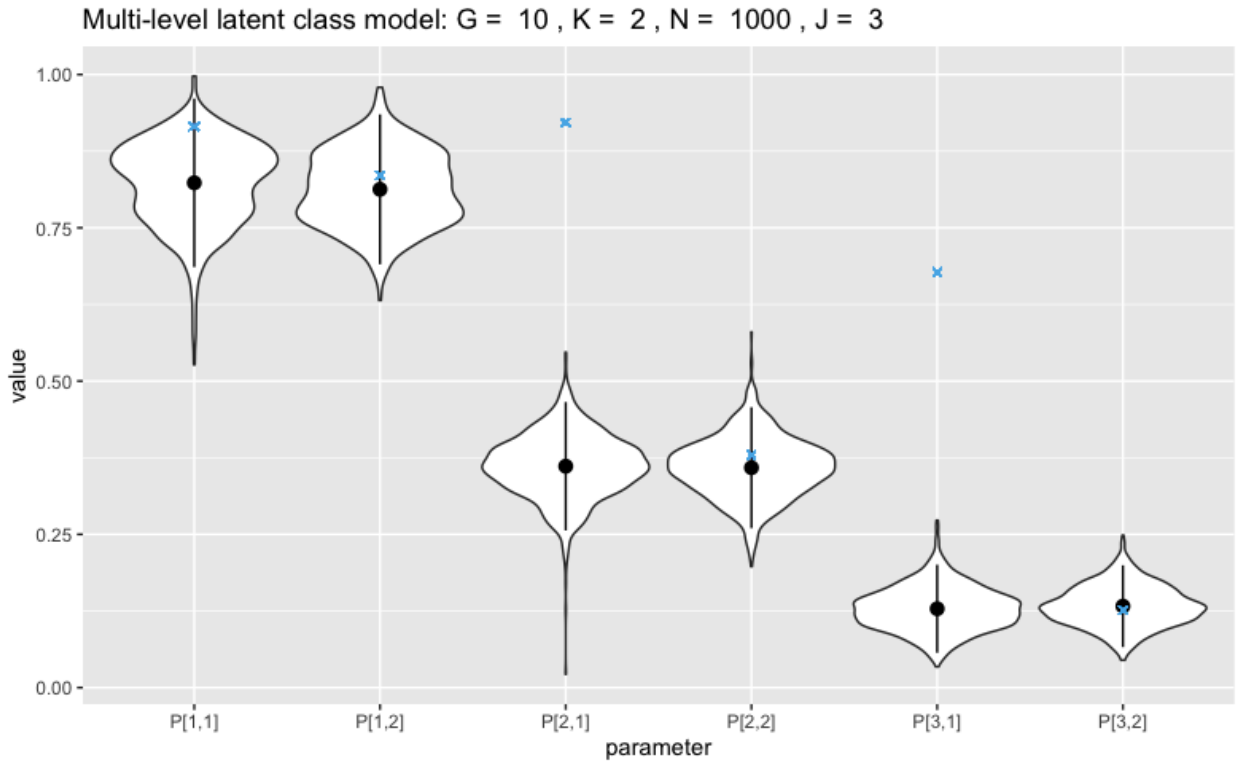
3. Examples and Evaluation

In the two simulated dataset with $G = 10, 100$, we let $K = 2$, $J = 3$ and $N = 1000$. In both settings, the estimation of PI seems to be strongly affected by prior distribution. What's more, the estimation of P deviated from prior distribution ($\beta(1, 1)$ with mean 0.5), they doesn't estimate the true value well.

3.1 Simulation Setting: G=10, J = 3, N = 1000, K = 2

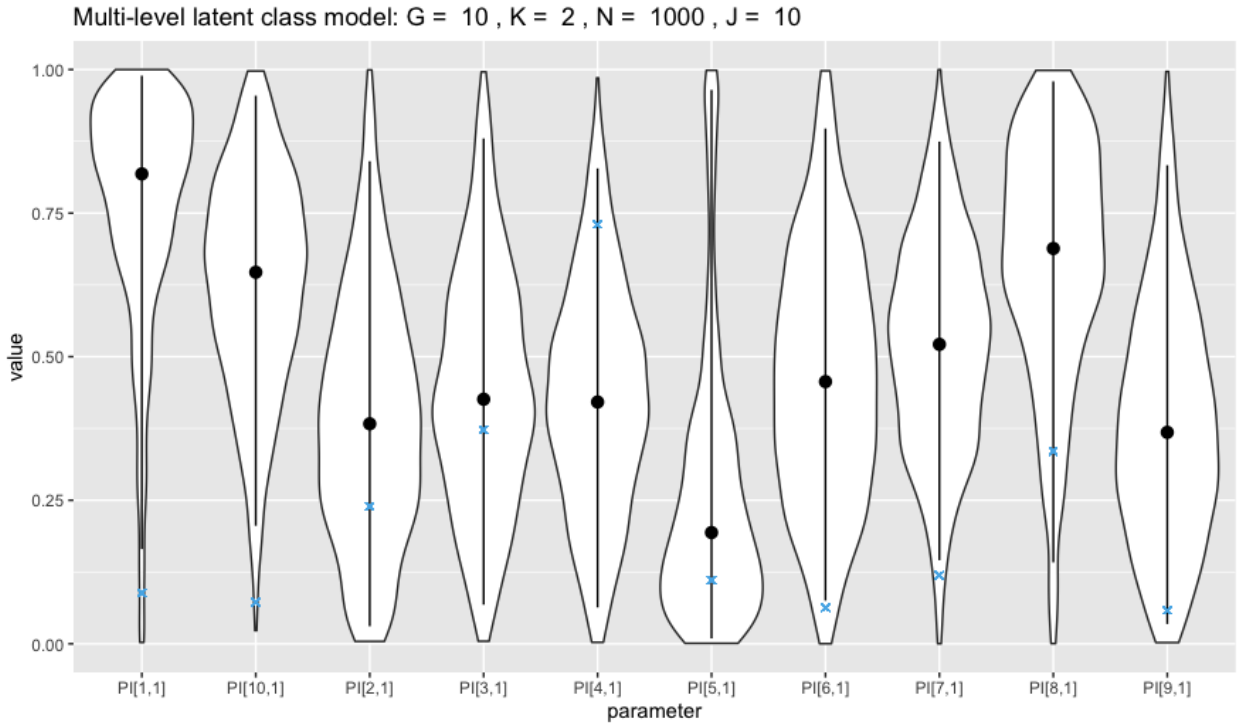
For both parameter *PI(G-K)* and *P(J-K)*, we use density plot to show the distribution of sampler. For latent class variable U, we use misclassfication rate to evaluate the model.

For P



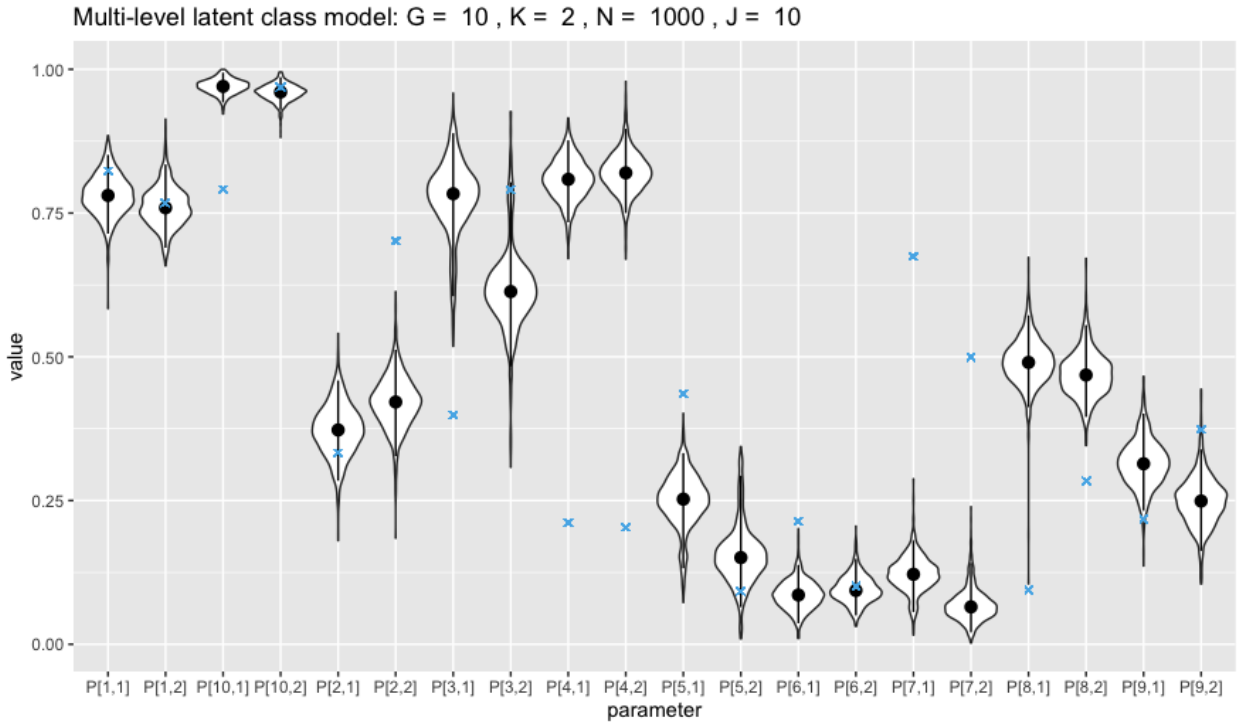
For PI

Since $\vec{\pi}^g$ has two dimension and they sum up to be 1, we only show the probability of latent class being 1.

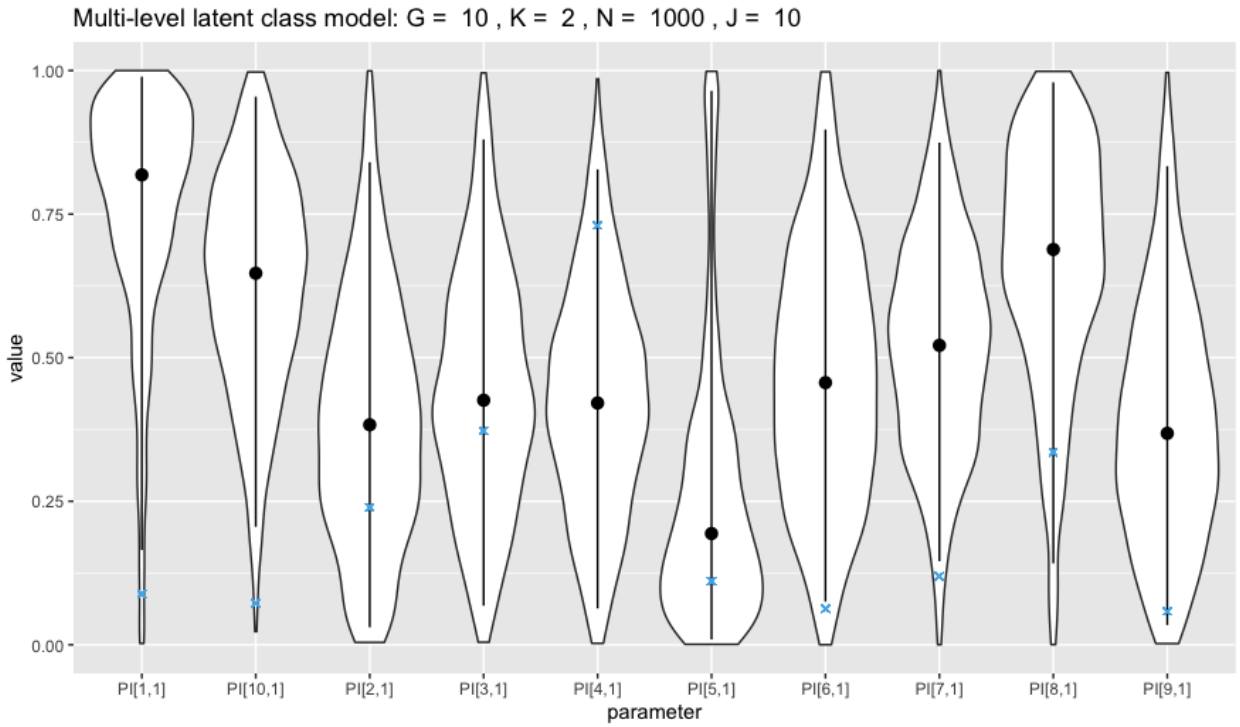


3.2 Simulation setting G = 10, J = 10, N = 1000, K = 2

For P



For PI



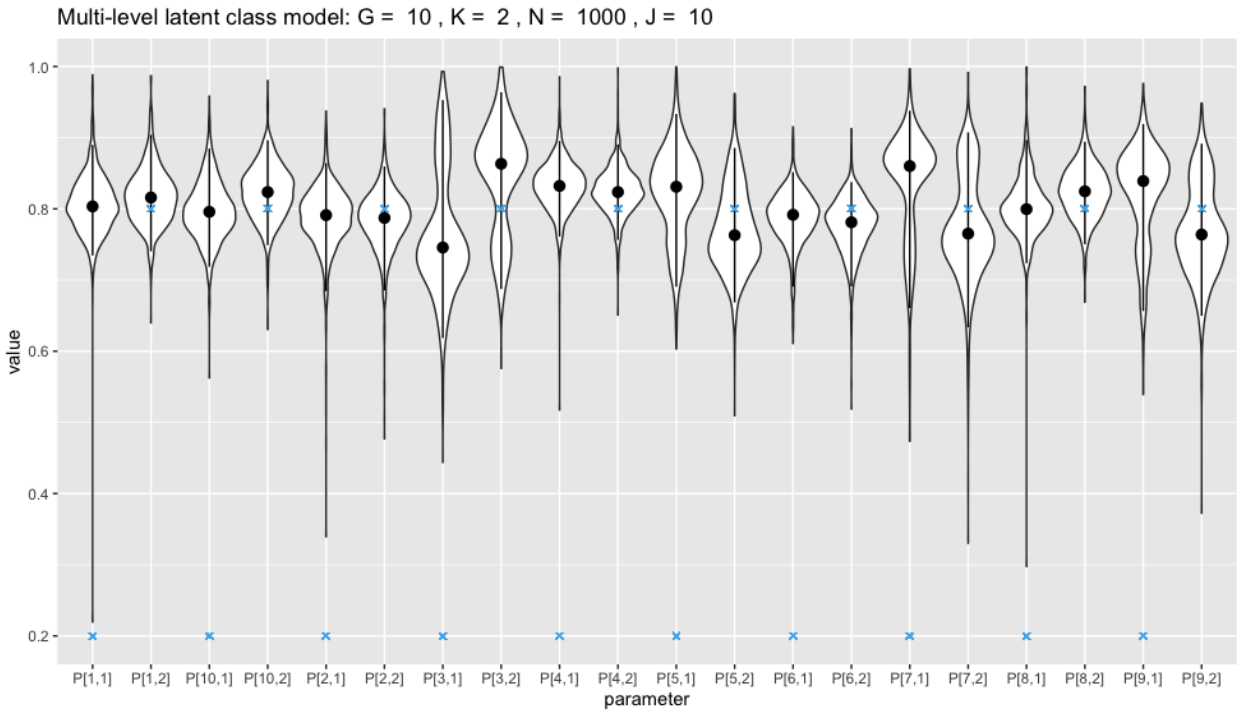
Conclusion: higher dimension of causes leads to more stable estimation (compared toJ = 3), however the estimation was not correct for some variable. So I need to check my model.

3.3 Simulation setting G = 10, J = 10, N = 1000, K = 2

This was a special setting when $A_j|U = 1 \sim Bernoulli(0.2), A_j|U = 2\ Bernoulli(0.8)$. Others remained the same part 3.2

For P

I have set the iteration time be 40,000 and burnout 30,000.



The traceplot of P is: