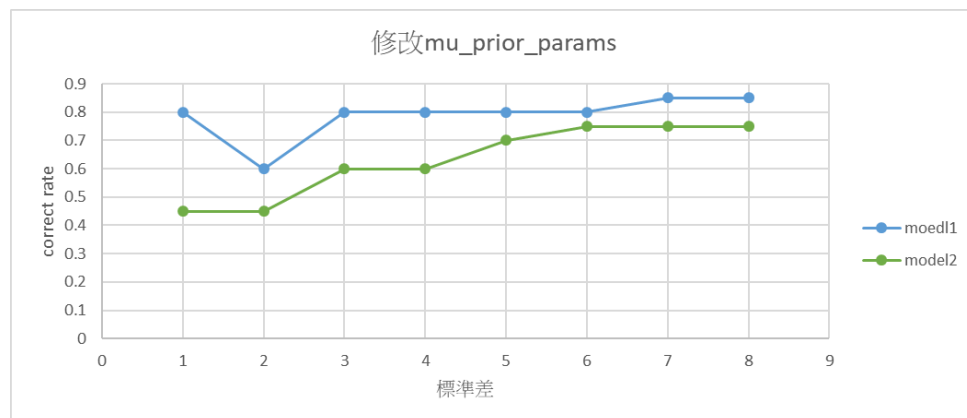
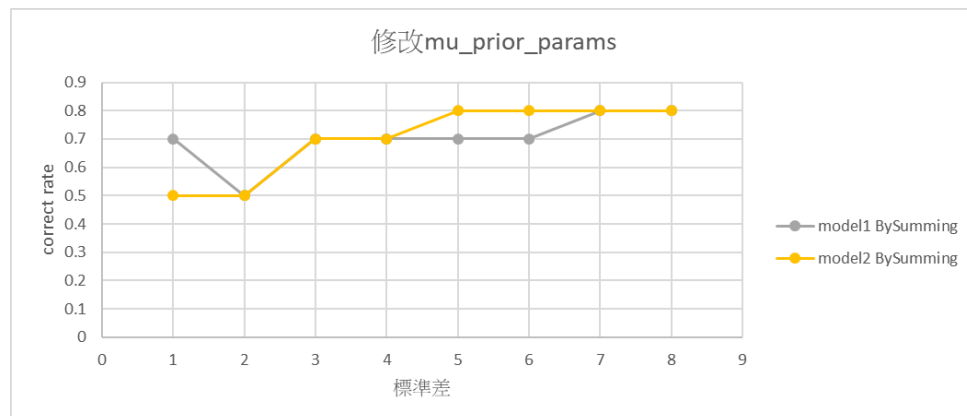
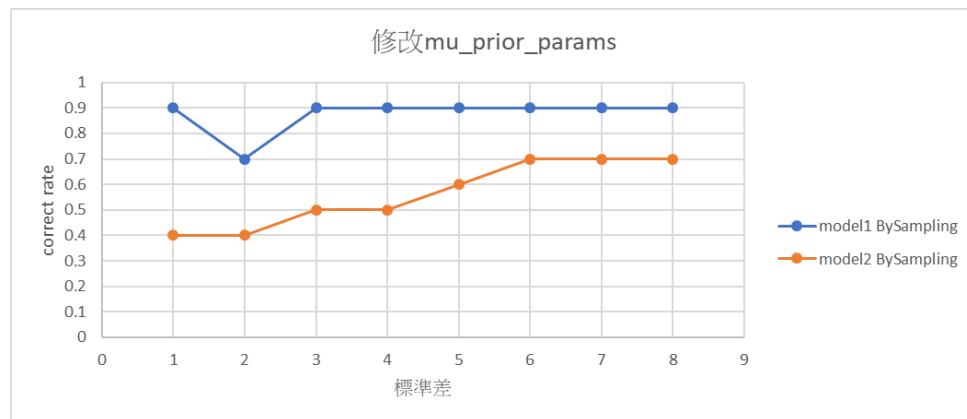


Different parameter settings

1. mu_prior_params - standard deviation

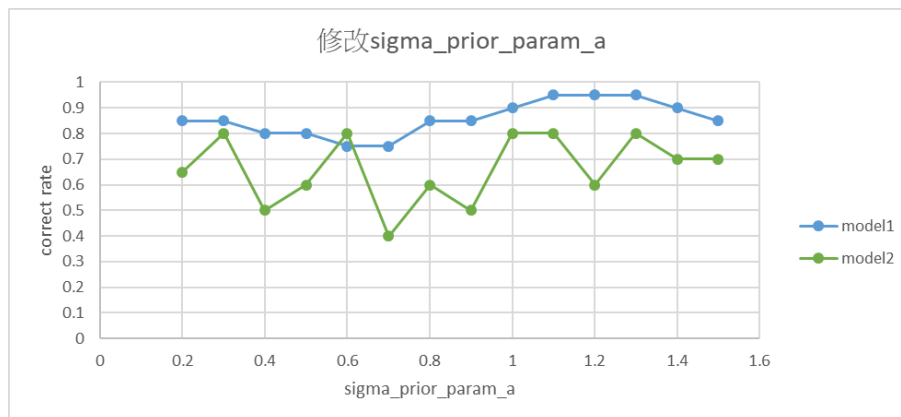
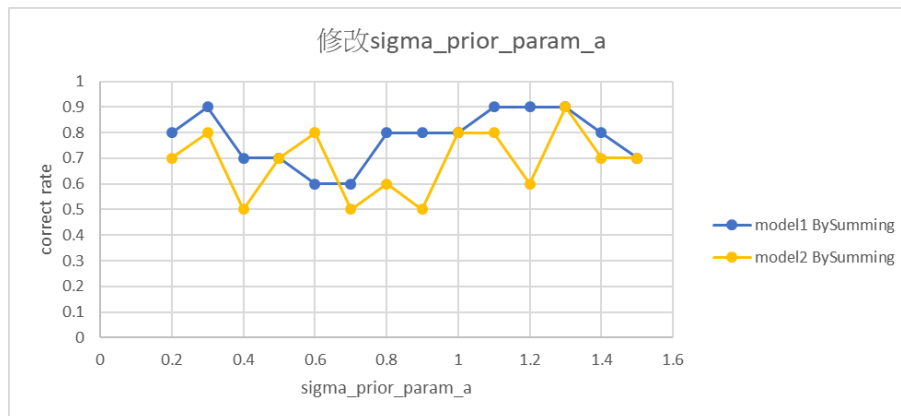
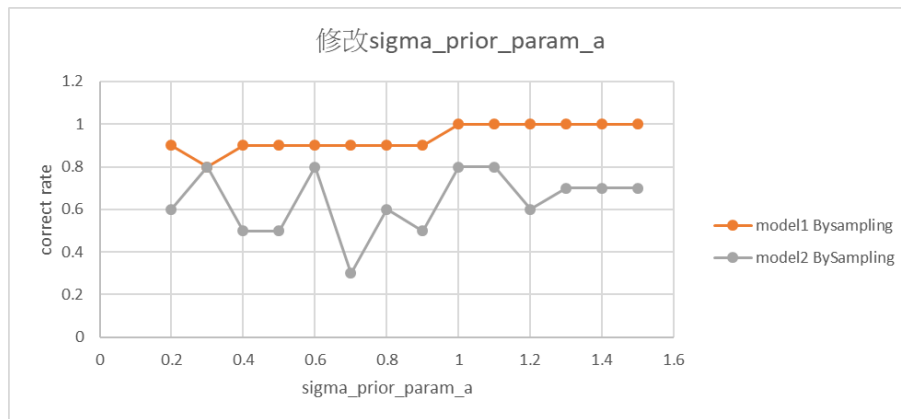


觀察：當 sigma 增加，正確率提升，且 model2 的提升比較明顯

推論：

- 較大的 sigma 可能會造成 mu 值的差異比較大，使得兩成分模型資料的分布更明顯分散，兩個成分之間的區隔更大，從而使得模型選擇算法更容易正確辨識出資料的混合成分結構。
- 當 sigma 較小時，兩成分的分布重疊度高，混合模型與單成分模型在數據表現上差異不大，導致辨識正確率較低。

2. sigma_prior_param_a



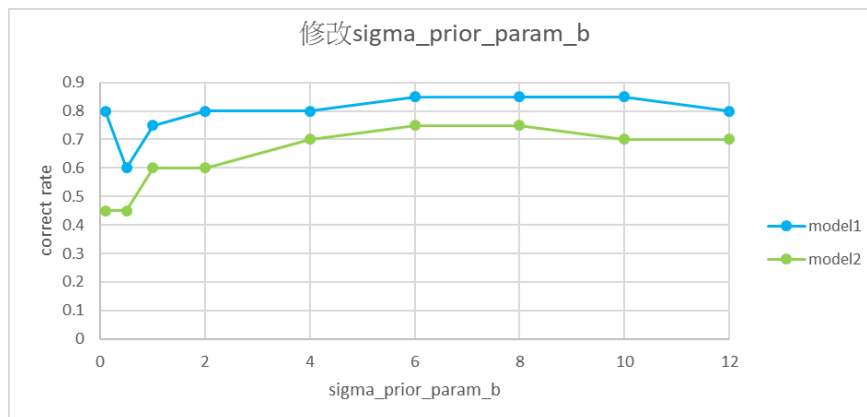
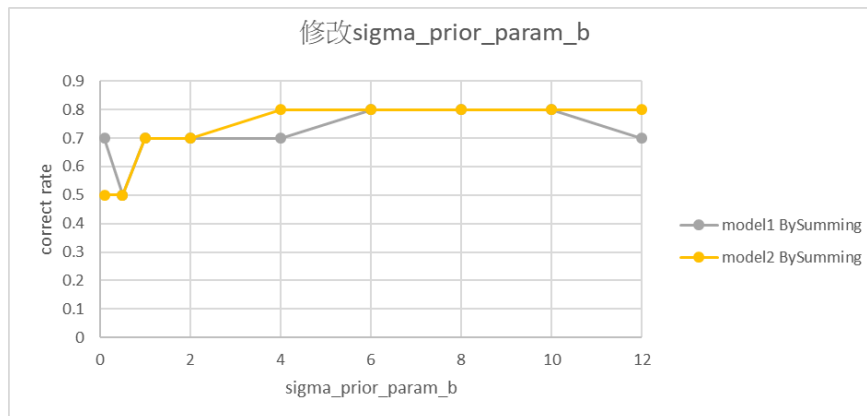
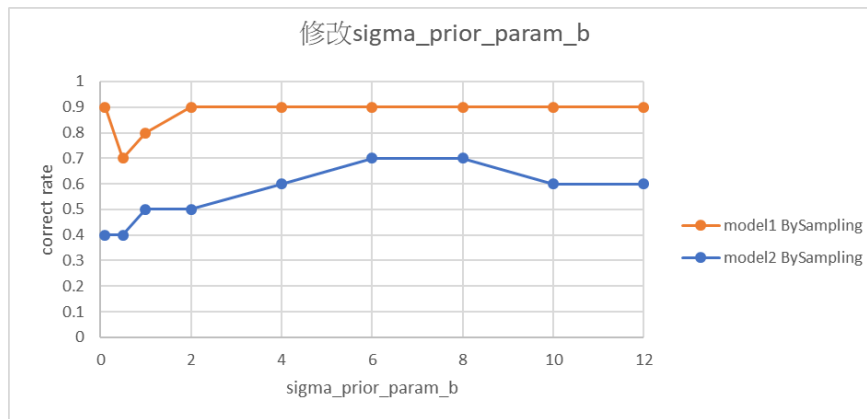
觀察：

model1 辨識較穩定，model2 較容易受 $\sigma_{\text{prior_param_a}}$ 的變動而波動

推論：

- model2 參數數量多，模型空間維度高，因此其邊際機率對於先驗分布的設定更為敏感。
- $\sigma_{\text{prior_param_a}}$ 決定了 Gamma 先驗分布形狀，進而影響對於 σ 的偏好。過小的 a 值高機率會有較大的 σ ，導致生成的樣本分布較扁平，造成混合模型與單一模型在數據表現上難以區分，模型 2 辨識準確率降低。

3. sigma_prior_param_b



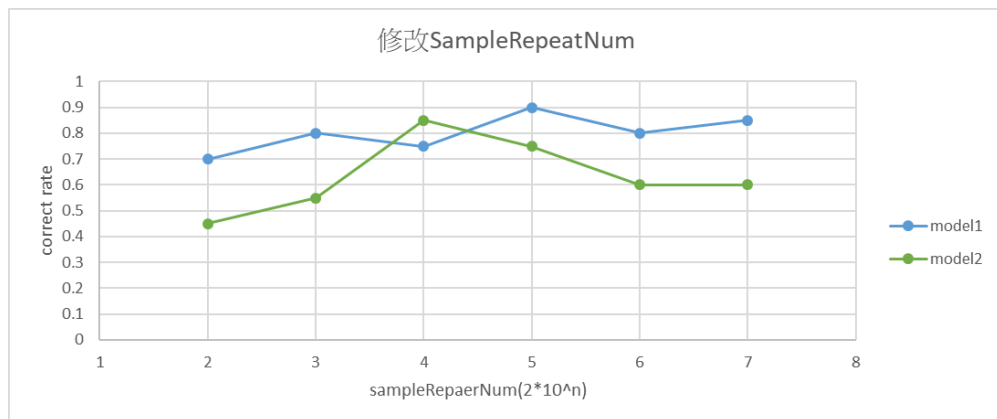
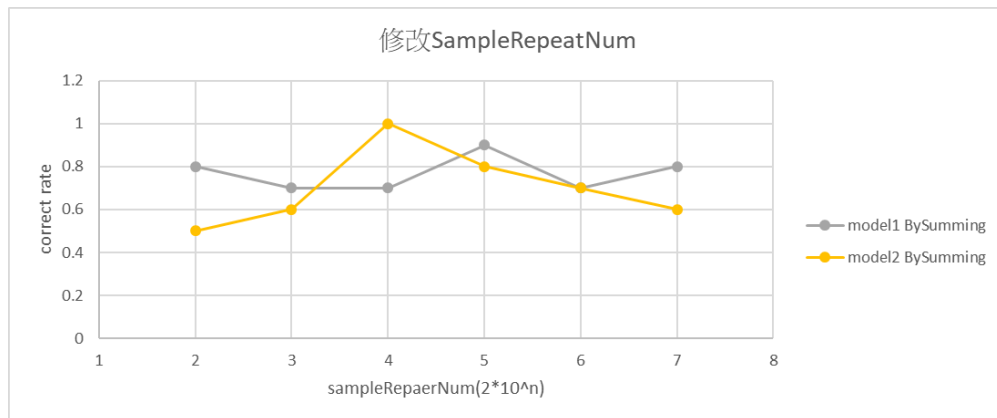
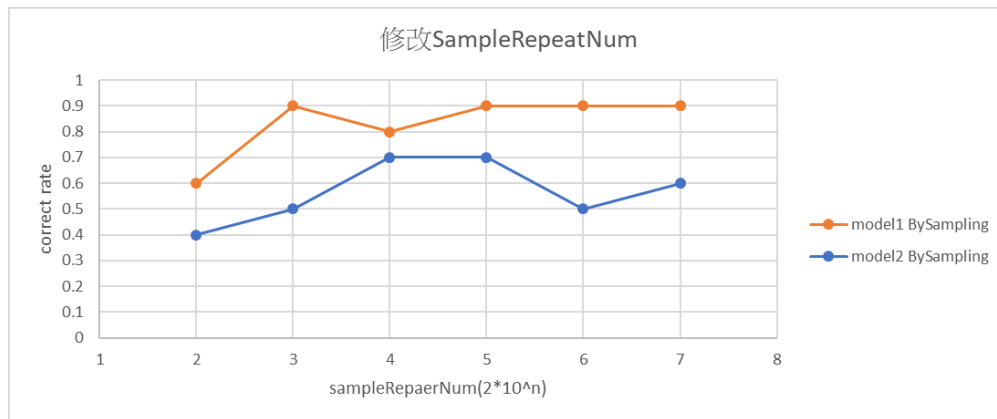
觀察：

- model1 辨識率始終高且穩定，幾乎不受 sigma_prior_param_b 改變影響；model2 辨識率隨 b 增大。當 b 值增加到一定程度時，結果趨於穩定

推論：

當 b 過小時，先驗會偏向尖銳分布，導致混合模型中兩個成分過度重疊、不易區分，使模型 2 辨識困難；隨著 b 增加，先驗更容許較寬的分布形狀，使兩個成分更能表現出差異，有助於模型 2 的辨識準確率提升。

4. sampleRepeatNum



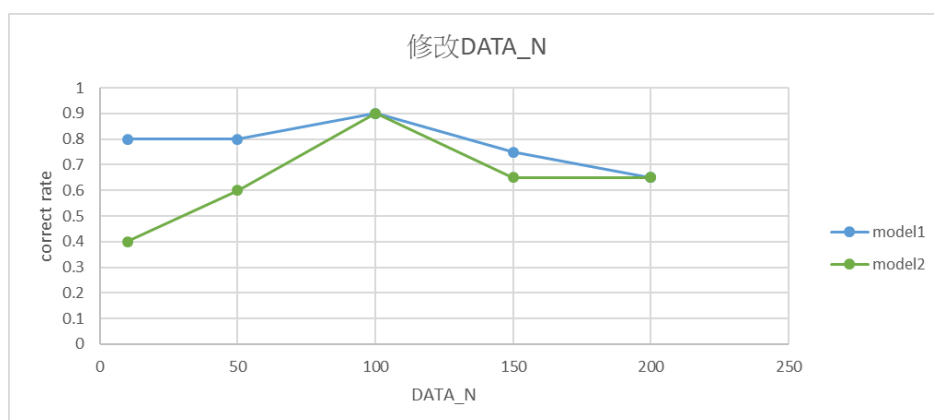
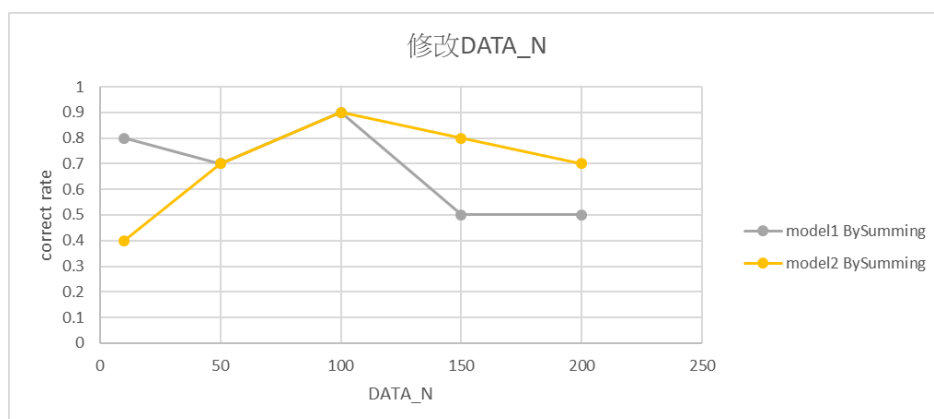
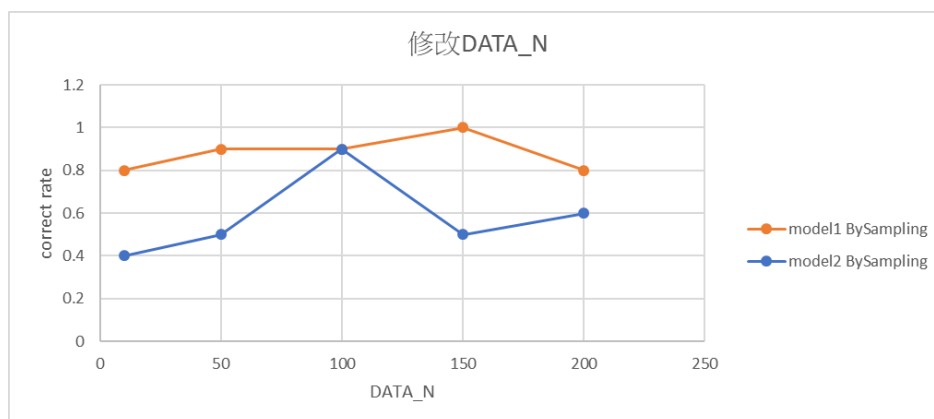
觀察：

- model1 只需中等規模的抽樣次數就能穩定辨識。
- model2 初期較依賴 sampleRepeatNum 的數量，但過大反而效果變差，呈現先升後降的趨勢。

推論：

- 對於複雜模型，在初期低抽樣次數時難以準確估計高維參數空間中的 evidence，因此正確率會隨 sampleRepeatNum 增加而提升。然而，當 sampleRepeatNum 過高時，可能導致數值積分震盪或估計誤差累積，反而使結果變差。

5. DATA_N



觀察：

- model1 辨識準確率較穩定；model2 波動較大
- 在 $N = 100$ 前有明顯增益，表示更多樣本能幫助區分成分，但再增加後卻效果下滑。

推論：

- 在貝式模型選擇中，樣本數對 evidence 的穩定性與代表性有重要影響。對 model1 而言，參數較少、結構簡單，即使在少量資料下，evidence 計算亦較穩定，辨識率高且波動小。

- 相對地，model2 需要辨識兩組參數與混合比例，因此在少量樣本下辨識困難。隨著樣本數增加，其模型結構才能充分反映於資料分布中，提升辨識準確率。
- 然而，當樣本數過多時（如 $N > 100$ ），反而可能出現 evidence 估算不穩、模型 overfit 或積分解析度不足的情況，使得 model2 辨識正確率下降。

Extra Credit

根據作業提供的相關公式來計算 2-component 的 Maximum likelihood，並且使用 Expectation-Maximization Algorithm 來實作

$$P[x_i|M_2, \mu_2, \sigma_2] = (1-b) \mathcal{N}(x_i; \mu_{2a}, \sigma_{2a}) + (b) \mathcal{N}(x_i; \mu_{2b}, \sigma_{2b})$$

參考：<https://chih-sheng-huang821.medium.com/機器學習-em-演算法-expectation-maximization-algorithm-em-高斯混合模型-gaussian-mixture-model-gmm-和 gmm-em 詳細推導-c6f634410483>

```
Data generated with two components
generating data with: m; ( $\mu_1, \sigma_1$ ); ( $\mu_2, \sigma_2$ ) = 0.078; (-1.11, 1.63); (-0.83, 3.15)
1-component maximum likelihood: 3.29161e-70
2-component mixture maximum likelihood: 8.63218e-55
Integrals by sampling= (6.588e-58, 9.03338e-58) by summing: (3.07579e-58, 8.55344e-58)
```

Write in Gaussian_pool0rNot1.c

Github: https://github.com/Yun0921/prob_model_class/tree/main/HW2