M526: Midterm

Dominic Bair

April 18, 2023

1. Here we create a Bayesian model for cluster analysis. We assume our measurements are generated from 4 clusters.

$$\begin{aligned}
\tau &\sim \text{Gamma}(A, B) \\
\tilde{\pi} &\sim \text{Dirichlet}_{\sigma_1,\sigma_2,\sigma_3,\sigma_4}(\alpha, \beta) \\
s_n|\tilde{\pi} &\sim \text{Categorical}_{\sigma_1,\sigma_2,\sigma_3,\sigma_4}(\tilde{\pi}), && n = 1, \cdots, N \\
w_n|s_n, \tau &\sim \text{Normal}\left(\mu_{s_n}, \frac{1}{\tau}\right), && n = 1, \cdots, N
\end{aligned}$$

   We choose $A = 2$, $B = 1/2$, $\alpha = 1$, $\beta = [1/4 \;\; 1/4 \;\; 1/4 \;\; 1/4]$. We choose these values based on a brief visual inspection of our data and our chosen distributions. See Figure 1.

2. Now we create a sampling scheme to approximate the posterior of our model. To achieve this, we implement a Gibbs sampler; which requires that we calculate the conditionals of each of our model parameters. We find the following conditionals:

$$\begin{aligned}
p(\tau|w_{1:N}) &= \text{Gamma}(\tau; A', B') \\
p(\tilde{\pi}|s_n) &= \text{Dirichlet}(\tilde{\pi}; \alpha\beta + \tilde{c}(s_n)) \\
p(s_n|\tau, s_{-n}, w_{1:N}) &= \text{Categorical}_{\sigma_1,\sigma_2,\sigma_3}(s_n; \tilde{\pi'}), && n = 1, \cdots, N.
\end{aligned}$$

   We have $A' = A + N$, $1/B' = 1/B + 1/2 \sum_{n=1}^{N} -(w_n - \mu_{s_n})^2$, and $\pi'_{\sigma_m} = \dfrac{h_m}{\sum_{j=1}^{M} h_j}$ with $h_m = \exp\left\{\tau/2(w_n - \mu_{s_n})^2\right\}$, for $m = 1, \cdots, M$.

3. We use these conditionals to create a Gibb's sampling scheme to sample from the posterior distribution of our model. See Figure 2.

4. Lastly, we use our Gibb's sampler to estimate the likelihood that w_{15}, and w_{25} are in the same cluster. We generate 1000 samples with our Gibb's sampler and find $p(s_{15} = s_{25}) = 0.0360$. See Figure 1 to see the location of w_{15} and w_25 in our data set.

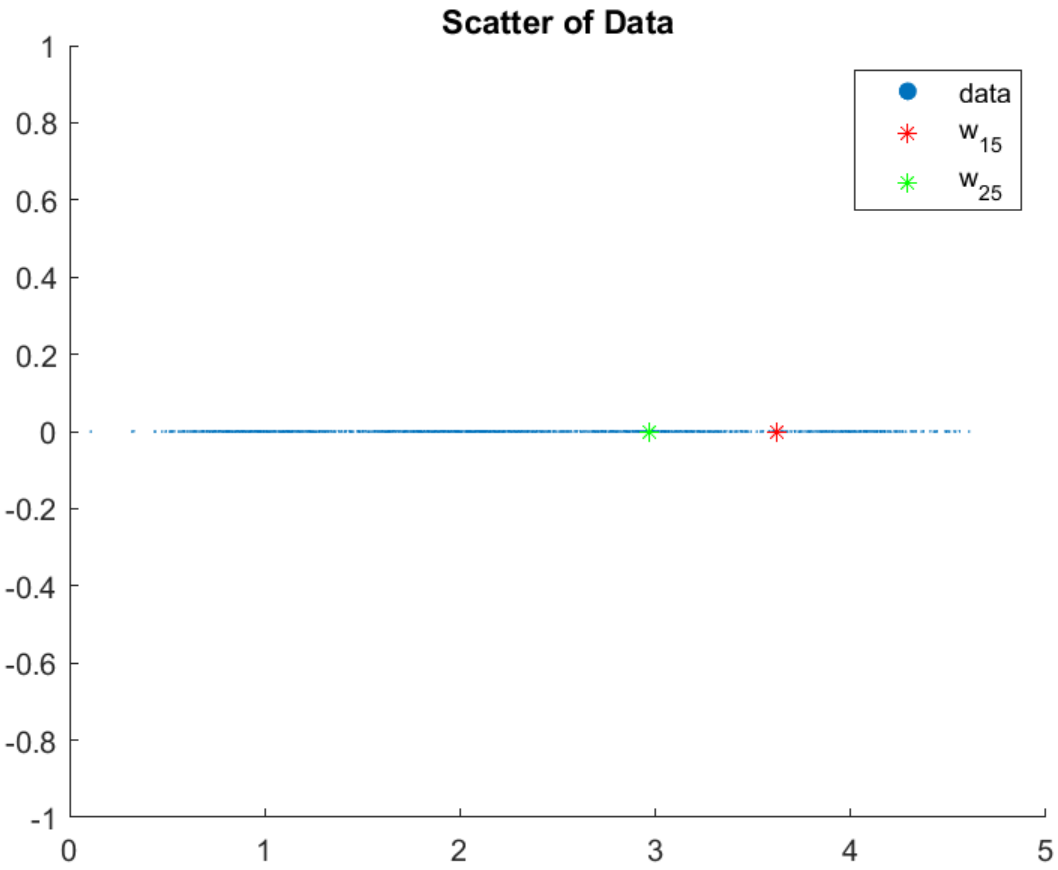

Figure 1: A scatter plot of our data. This visualization is primarily used to show w_{15}, and w_{25} in relation to one another.

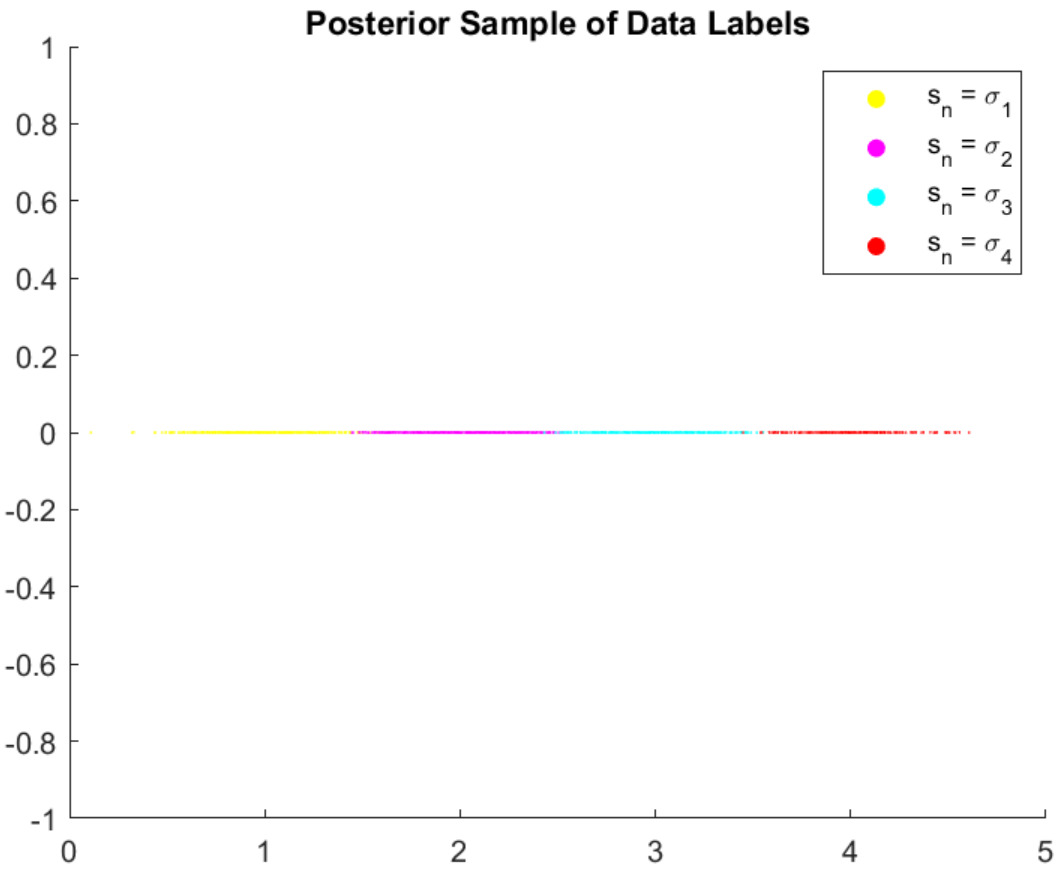

Figure 2: Here we plot one sample of $s_{1:N}$ from our modeled posterior. Each s_n is a label for its associated w_n. We see our Gibb's sampler separates our data in an expected way with little overlap in the clusters.