M526: HW2

Dominic Bair

March 21, 2023

1. Here we create a Bayesian model for cluster analysis. We assume our measurements are generated from 3 clusters

$$\begin{aligned}
X_{\sigma_m} &\sim \text{Normal}(R_x, C_x), && m = 1,2,3 \\
Y_{\sigma_m} &\sim \text{Normal}(R_y, C_y), && m = 1,2,3 \\
\tau &\sim \text{Gamma}(A, B) \\
\tilde{\pi} &\sim \text{Dirichlet}_{\sigma_1,\sigma_2,\sigma_3}(\alpha, \beta) \\
s_n|\tilde{\pi} &\sim \text{Categorical}_{\sigma_1,\sigma_2,\sigma_3}(\tilde{\pi}), && n = 1, \cdots, N \\
x_n|s_n, X_{\sigma_1}, X_{\sigma_2}, X_{\sigma_3}, \tau &\sim \text{Normal}\left(X_{s_n}, \frac{1}{\tau}\right), && n = 1, \cdots, N \\
y_n|s_n, Y_{\sigma_1}, Y_{\sigma_2}, Y_{\sigma_3}, \tau &\sim \text{Normal}\left(Y_{s_n}, \frac{1}{\tau}\right), && n = 1, \cdots, N
\end{aligned}$$

   We assume our measurements are generated from 3 clusters; thus, each data point belongs to one of our 3 clusters, σ_1, σ_2, or σ_3. We choose s_n to be the categorical random variable that describes this. We choose $\tilde{\pi}$ to be a Dirichlet random variable since each point must belong to a category. We choose τ to be a gamma random variable since we cannot have negative variances. Lastly, we choose X_{σ_m} and Y_{σ_m}, $m = 1, 2, 3$ to be normal random variables because we expect each cluster to be roughly normal.

   We choose our hyper parameters as follows: $R_x = [-3, 3, 7]$, $R_y = [3, 0, 6]$, $C_x = C_y = 1$, $A = 2$, $B = 1/2$, $\alpha = 1$, and $\beta = [1/3, 1/3, 1/3]$. We choose these values based on a brief visual inspection of our data and our chosen distributions.

2. Next, we create a graphical representation of our model. See Figure 1.

3. 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.

   First, we find the conditional for $\tilde{\pi}$ to be

$$p(\tilde{\pi}|s_n) = \text{Dirichlet}(\tilde{\pi}; \alpha\beta = \tilde{c}(s_n)).$$

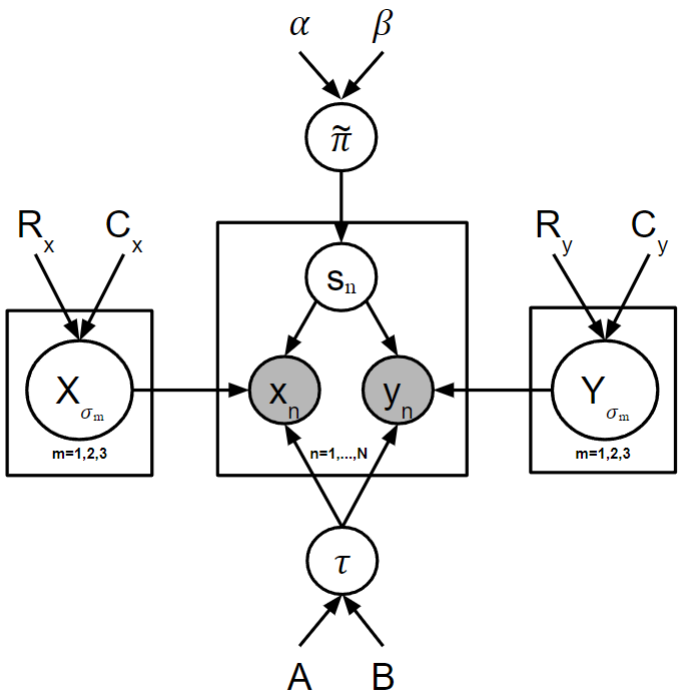

Figure 1: Graphical representation of our clustering model.

Nest, we find the conditional for s_n to be

$$p(s_n|\tau, s_{-n}, x_{1:N}, y_{1:N}, s_{1:N}, X_{\sigma_{1:M}}, Y_{\sigma_{1:M}}) = \text{Categorical}_{\sigma_1,\sigma_2,\sigma_3}(s_n; \tilde{\pi'}), \quad n = 1, \cdots, N.$$

Where

$$\tilde{\pi'}(m) \propto \exp\left\{(x_n - X_{\sigma_1})^2 + (y_n - Y_{\sigma_1})^2\right\} \qquad m = 1, 2, 3.$$

Now we find the conditional for X_{σ_m} to be

$$p(X_{\sigma_m}|\tau, X_{\sigma\_m}, x_{1:N}, y_{1:N}, s_1 : N) = \text{Normal}(X_{\sigma_m}; \mu, v), \qquad m = 1, 2, 3.$$

Where

$$\mu = \frac{C_x\tau\sum_{n=1}^{N} x_n + R_x}{C_x\tau N + 1}, \text{ and } v = \frac{1}{\tau N + 1/C_x}.$$

Similarly we find the conditional for Y_{σ_m} to be

$$p(Y_{\sigma_m}|\tau, Y_{\sigma\_m}, x_{1:N}, y_{1:N}, s_1 : N) = \text{Normal}(Y_{\sigma_m}; \mu, v), \qquad m = 1, 2, 3.$$

Where

$$\mu = \frac{C_y\tau\sum_{n=1}^{N} y_n + R_y}{C_y\tau N + 1}, \text{ and } v = \frac{1}{\tau N + 1/C_y}.$$

Finally, we find the conditional for τ to be

$$p(\tau|x_{1:N}, y_{1:N}, s_{1:N}, X_{\sigma_{1:M}}, Y_{\sigma_{1:M}}) = \text{Gamma}(\tau; A', B').$$

Where

$$A' = A + N, \text{ and } 1/B' = 1/B + 1/2 \sum_{n=1}^{N} (x_n - X_{s_n})^2 + (y_n - Y_{s_n})^2, \qquad n = 1, \cdots, N.$$

Now that we have our conditionals for all variables, we pick initial values for all variables and proceed to generate new values for each of them via a Gibbs sampling scheme. We begin by a generating $\tilde{\pi}$ based on its conditional, and our initial value. Then, we generate N, s_n values based on their conditionals and our initial values. We repeat this process until we have generated one new value for each of our model parameters. Lastly, we repeat this process J times so we have an approximation of our posterior distribution.

4. We pick $J = 1000$ and execute our Gibbs sampler. We plot all of our generated X_{σ_m} and Y_{σ_m}. See Figure 2.

5. Finally, we use our approximation of our posterior to estimate the probability that our 11th data point and 982nd data point are in the same cluster. We do this by simply counting the number of times s_{11} is equal to s_{982} from our generated samples and divide by J. We find the probability to be 0.8360. Thus, our 982nd datum point likely belongs to the second cluster. See Figure 2.

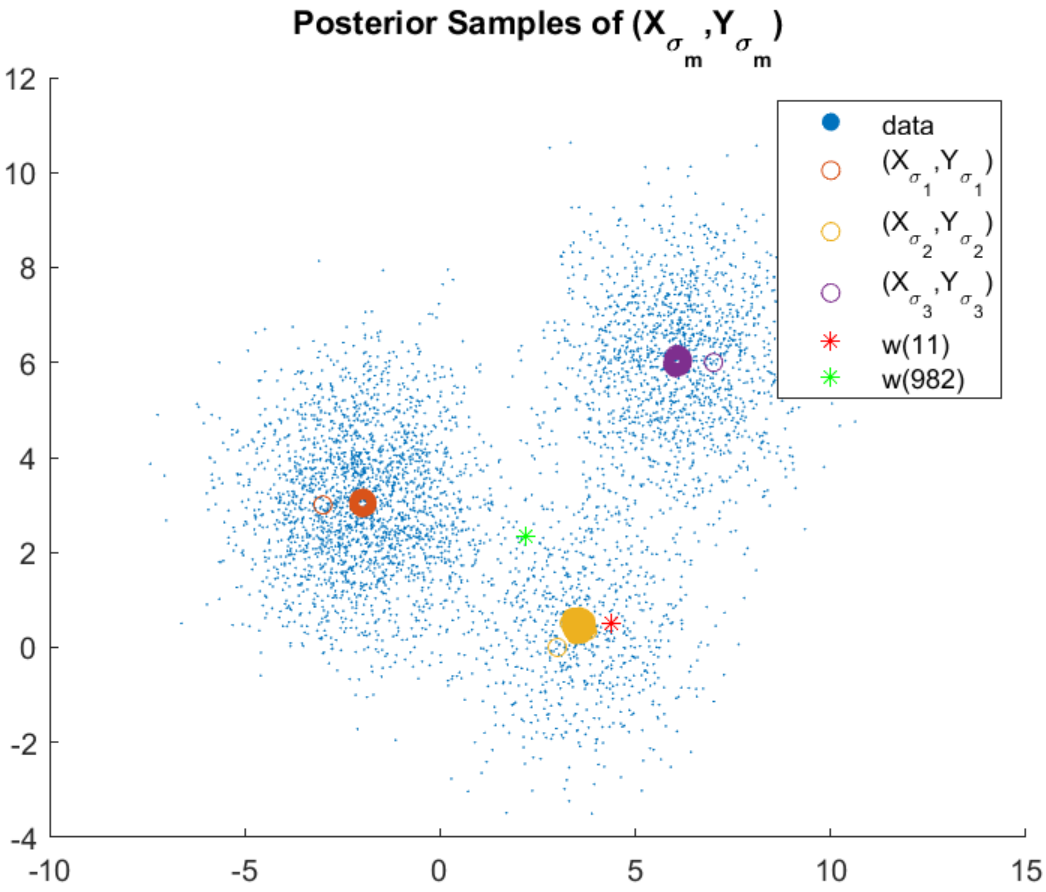

Figure 2: A plot of our given data, generated values of X_{σ_m} and Y_{σ_m}, with our 11th and 982nd data points highlighted. Note that our generated cluster centers are very tightly clustered with the only outliers being our initial values. We clearly see that the 11th datum point belongs in the second cluster, but the 982nd datum point may belong to a different cluster.