STATS 551 Homework 3

..

2023-10-09

Problem 3

(a)

By definition of this mixture distribution, we have

$$
\begin{aligned}
&p(Z_i = j) = p_j,\ j = 1, 2, \quad Z_i \text{ i.i.d} \\
&X_i | Z_i = 1 \sim N(\mu_1, \Sigma_1), \quad X_i | Z_i = 2 \sim N(\mu_2, \Sigma_2), \quad X_i \text{ i.i.d}
\end{aligned}
$$

It's worth noting that Z_i follows a discrete distribution with only two parameters and we have $p_1 + p_2 = 1$, so it's more convenient to express the distribution as Bernoulli, i.e.

$$
C_i = (Z_i - 1) | p_1 \overset{\text{i.i.d}}{\sim} \text{Bern}(p_2),\ \text{i.e. } p(C_i = 1) = p_2
$$

So, in the following, we use C_i, p_2, which is equavilent to Z_i, p_1, p_2

We need to specify priors for $p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2$. Note that these parameters exhibit certain structures, i.e. we should assume the following independencies

$$
p(p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2) = p(p_2) \cdot p(\mu_1, \Sigma_1) \cdot p(\mu_2, \Sigma_2)
$$

This assumption is natural, since in a mixture model, different components should not be interrelated, and the same holds between components and class labels.

For simplicity, we use beta distribution for $p(p_2)$ and Jeffreys' prior for $p(\mu_1, \Sigma_1)$ and $p(\mu_2, \Sigma_2)$. That is

$$
\begin{aligned}
p(p_2) &\sim \text{Beta}(\alpha, \beta) \\
p(\mu_1, \Sigma_1) &\propto |\Sigma_1|^{-5/2} \\
p(\mu_2, \Sigma_2) &\propto |\Sigma_2|^{-5/2}
\end{aligned}
$$

For hyperparameter (α, β), we didn't get any extra information about the clusters, so we just choose $\alpha = \beta = 1$.

Overall, the model is

$$
\begin{aligned}
&p(p_2) \sim \text{Beta}(\alpha, \beta) \\
&p(\mu_1, \Sigma_1) \propto |\Sigma_1|^{-5/2} \\
&p(\mu_2, \Sigma_2) \propto |\Sigma_2|^{-5/2} \\
&p(p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2) = p(p_2) \cdot p(\mu_1, \Sigma_1) \cdot p(\mu_2, \Sigma_2) \\
&C_i | p_2 \overset{\text{i.i.d}}{\sim} \text{Bern}(p_2) \\
&X_i | C_i = 0 \sim N(\mu_1, \Sigma_1),\ X_i | C_i = 1 \sim N(\mu_2, \Sigma_2),\ X_i \text{ i.i.d}
\end{aligned}
$$

(b)

Note that only X_is are observable and it's helpful to introduce C_is to the posterior.
Joint posterior

$$
\begin{aligned}
&p(p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2, C_1, \cdots, C_n|\mathbf{X}) \\
\propto\, & p(\mathbf{X}|p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2, C_1, \cdots, C_n) \cdot p(p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2, C_1, \cdots, C_n) \\
\propto\, & \prod_i p(X_i|C_i, \mu_1, \Sigma_1, \mu_2, \Sigma_2) \cdot p(C_1, \cdots, C_n|p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2) \cdot p(p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2) \\
\propto\, & \left(\prod_i \Big[(1-C_i)N(X_i|\mu_1, \Sigma_1) + C_i N(X_i|\mu_2, \Sigma_2)\Big]\right) \cdot \left(\prod_i p(C_i|p_2)\right) \cdot p(p_2) \cdot p(\mu_1, \Sigma_1) \cdot p(\mu_2, \Sigma_2)
\end{aligned}
$$

Now we consider the full conditional distributions

$$
\begin{aligned}
p(p_2|\cdot) &\propto \left(\prod_i p(C_i|p_2)\right) \cdot p(p_2) \\
&\propto \prod_i (1-p_2)^{1-C_i} p_2^{C_i} \\
&\propto p_2^{n-\Sigma C_i}(1-p_2)^{\Sigma C_i} \\
&\sim \text{Beta}(n - \Sigma C_i + 1, \Sigma C_i + 1)
\end{aligned}
$$

$$
p(\mu_1|\cdot) \propto \left(\prod_i \Big[(1-C_i)N(X_i|\mu_1, \Sigma_1) + C_i N(X_i|\mu_2, \Sigma_2)\Big]\right) \cdot p(\mu_1, \Sigma_1)
$$

Note that C_i is a binary variable, so it's convenient to rewrite

$$
\begin{aligned}
p(\mu_1|\cdot) &\propto \left(\prod_i \left|\Sigma_i^{\text{mix}}\right|^{-1/2}\right) \exp\left[-\frac{1}{2} \sum_{i:\, C_i=0} (X_i - \mu_1)^T \Sigma_1^{-1} (X_i - \mu_1)\right] \\
&\qquad \times \exp\left[-\frac{1}{2} \sum_{i:\, C_i=1} (X_i - \mu_2)^T \Sigma_2^{-1} (X_i - \mu_2)\right] \\
&\propto \exp\left[-\frac{1}{2} \sum_{i:\, C_i=0} (\mu_1 - X_i)^T \Sigma_1^{-1} (\mu_1 - X_i)\right] \\
&\propto \exp\left[-\frac{1}{2}\Big(\text{tr}\left(\Sigma_1^{-1} S_1\right) + n_1(\mu_1 - \bar{X}_1)^T \Sigma_1^{-1} (\mu_1 - \bar{X}_1)\Big)\right] \\
&\sim N(\bar{X}_1, \frac{\Sigma_1}{n_1})
\end{aligned}
$$

in which

$$
\begin{aligned}
\Sigma_i^{\text{mix}} &= (1-C_i)\Sigma_1 + C_i\Sigma_2 \\
n_1 &= n - \sum_i C_i \\
\bar{X}_1 &= \frac{1}{n_1} \sum_{i:\, C_i=0} X_i \\
S_1 &= \sum_{i:\, C_i=0}^{n} (X_i - \bar{X}_1)(X_i - \bar{X}_1)^T
\end{aligned}
$$

Similarly, we can get

$$
p(\mu_2|\cdot) \sim N(\bar{X}_2, \frac{\Sigma_2}{n_2})
$$

in which

$$n_2 = \sum_i C_i$$

$$\bar{X}_2 = \frac{1}{n_2} \sum_{i:\, C_i=1} X_i$$

For Σ_1 and Σ_2, we have

$$\begin{aligned}
p(\Sigma_1|\cdot) &\propto \left(\prod_i \Big[(1-C_i)N(X_i|\mu_1,\Sigma_1) + C_i N(X_i|\mu_2,\Sigma_2)\Big]\right) \cdot p(\mu_1,\Sigma_1) \\
&\propto \left(\prod_i \left|\Sigma_i^{\text{mix}}\right|^{-1/2}\right) \exp\left[-\frac{1}{2} \sum_{i:\, C_i=0} (X_i-\mu_1)^T \Sigma_1^{-1} (X_i-\mu_1)\right] |\Sigma_1|^{-5/2} \\
&\propto |\Sigma_1|^{-n_1/2} |\Sigma_1|^{-5/2} \exp\left[-\frac{1}{2} \sum_{i:\, C_i=0} (X_i-\mu_1)^T \Sigma_1^{-1} (X_i-\mu_1)\right] \\
&\propto |\Sigma_1|^{-(n_1+1+3+1)/2} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\Sigma_1^{-1} S_1^0\right)\right] \\
&\propto \text{Inv-Wishart}_{n_1+1}((S_1^0)^{-1})
\end{aligned}$$

in which n_1 follows above, and

$$S_1^0 = \sum_{i:\, C_i=0} (X_i-\mu_1)(X_i-\mu_1)^T$$

Similarly,

$$p(\Sigma_2|\cdot) \sim \text{Inv-Wishart}_{n_2+1}((S_2^0)^{-1})$$

in which n_2 follows above, and

$$S_1^0 = \sum_{i:\, C_i=1} (X_i-\mu_2)(X_i-\mu_2)^T$$

For C_is,

$$\begin{aligned}
p(C_i|\cdot) &\propto \left(\prod_i \Big[(1-C_i)N(X_i|\mu_1,\Sigma_1) + C_i N(X_i|\mu_2,\Sigma_2)\Big]\right) \cdot \left(\prod_i p(C_i|p_2)\right) \\
&\propto \Big[(1-C_i)N(X_i|\mu_1,\Sigma_1) + C_i N(X_i|\mu_2,\Sigma_2)\Big] p_2^{C_i}(1-p_2)^{1-C_i}
\end{aligned}$$

in which, we use $N(X_i,|\cdot,\cdot)$ to denote the density function of multivariate normal distribution.

Overall, we have

$$\begin{aligned}
&p(p_2|\cdot) \sim \text{Beta}(n_1+1, n_2+1) \\
&p(\mu_1|\cdot) \sim N(\bar{X}_1, \frac{\Sigma_1}{n_1}) \\
&p(\mu_2|\cdot) \sim N(\bar{X}_2, \frac{\Sigma_2}{n_2}) \\
&p(\Sigma_1|\cdot) \sim \text{Inv-Wishart}_{n_1+1}((S_1^0)^{-1}) \\
&p(\Sigma_2|\cdot) \sim \text{Inv-Wishart}_{n_2+1}((S_2^0)^{-1}) \\
&p(C_i|\cdot) \propto \Big[(1-C_i)N(X_i|\mu_1,\Sigma_1) + C_i N(X_i|\mu_2,\Sigma_2)\Big] p_2^{C_i}(1-p_2)^{1-C_i}
\end{aligned}$$

(c)

We draw samples for $(p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2, C_1, \cdots, C_n)$. As for $p_1, Z_1, \cdots, Z_n$, just use

$$p_1 = 1 - p_2,\;\; Z_i = C_i + 1$$

```r
mixgauss <- read.table("./mixgauss.dat", header=FALSE)
```

Problem 4

(a)

We consider the normal model of multiple observations. The likelihood is

$$
\begin{aligned}
p(\mathbf{y}|\theta,\sigma^2) &= \prod_i p(y_i|\theta,\sigma^2) \\
&= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right]
\end{aligned}
$$

so

$$
\begin{aligned}
\log p(\mathbf{y}|\theta,\sigma^2) &= -\sum_i \log\sqrt{2\pi\sigma^2} + \frac{(y_i-\mu)^2}{2\sigma^2} \\
&= \text{const} - \frac{n}{2}\log\sigma^2 - \frac{(n-1)s_y^2 + n(\bar{y}-\mu)^2}{2\sigma^2}
\end{aligned}
$$

in which, $s_y^2 = \sum_i (y_i-\bar{y})^2/(n-1)$.
Therefore, we have

$$
\begin{aligned}
\frac{\partial \log p(\mathbf{y}|\theta,\sigma^2)}{\partial\mu} &= \frac{n(\bar{y}-\mu)}{\sigma^2} \\
\frac{\partial \log p(\mathbf{y}|\theta,\sigma^2)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{(n-1)s_y^2 + n(\bar{y}-\mu)^2}{2(\sigma^2)^2}
\end{aligned}
$$

and then

$$
\begin{aligned}
\frac{\partial^2 \log p(\mathbf{y}|\theta,\sigma^2)}{\partial\mu^2} &= -\frac{n}{\sigma^2} \\
\frac{\partial^2 \log p(\mathbf{y}|\theta,\sigma^2)}{\partial\sigma^2} &= \frac{n}{2\sigma^4} - \frac{(n-1)s_y^2 + n(\bar{y}-\mu)^2}{\sigma^6} \\
\frac{\partial^2 \log p(\mathbf{y}|\theta,\sigma^2)}{\partial\mu\partial\sigma^2} &= \frac{\partial^2 \log p(\mathbf{y}|\theta,\sigma^2)}{\partial\sigma^2\partial\mu} = -\frac{n(\bar{y}-\mu)}{\sigma^4}
\end{aligned}
$$

Therefore, the Fisher Information is

$$
\begin{aligned}
I(\mu,\sigma^2) &= -\mathbb{E}\begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{n(\bar{y}-\mu)}{\sigma^4} \\ -\frac{n(\bar{y}-\mu)}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{(n-1)s_y^2+n(\bar{y}-\mu)^2}{\sigma^6} \end{bmatrix} \\
&= -\begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} - \frac{(n-1)\sigma^2+\sigma^2}{\sigma^6} \end{bmatrix} \\
&= \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}
\end{aligned}
$$

in which, we use

$$
\mathbb{E}(\bar{y}-\mu) = 0, \mathbb{E}(\bar{y}-\mu)^2 = \mathbb{E}s_y^2 = \sigma^2
$$

Therefore, the Jeffreys' prior is

$$
p_J(\mu,\sigma^2) \propto \sqrt{\frac{n^2}{2\sigma^6}} \propto (\sigma^2)^{-3/2}
$$

(b)

The posterior distribution is

$$\begin{aligned}
p_J(\mu,\sigma^2|\mathbf{y}) &\propto p_J(\mu,\sigma^2)p(\mathbf{y}|\mu,\sigma^2) \\
&\propto (\sigma^2)^{-3/2}\prod_i \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] \\
&\propto \left(\sigma^2\right)^{-3/2}\cdot\left(\sigma^2\right)^{-n/2}\exp\left[-\frac{(n-1)s_y^2+n(\bar{y}-\mu)^2}{2\sigma^2}\right] \\
&\propto \sigma^{-1}\cdot(\sigma^2)^{-(n/2+1)}\exp\left[-\frac{1}{2\sigma^2}\left(n\cdot\frac{\sum_i(y_i-\bar{y})^2}{n}+n(\bar{y}-\mu)^2\right)\right]
\end{aligned}$$

It's known that this term follows a Normal-Inverse-χ^2 distribution, formally (using the notation in the book Bayesian Data Analysis Third edition),

$$p_J(\mu,\sigma^2|\mathbf{y}) \sim \text{N-Inv-}\chi^2\left(\bar{y},\frac{\sum_i(y_i-\bar{y})^2}{n^2};n,\frac{\sum_i(y_i-\bar{y})^2}{n}\right)$$

To see more clearly, we can rewrite

$$\begin{aligned}
p_J(\mu|\sigma^2,\mathbf{y}) &\propto \sigma^{-1}\exp\left[-\frac{1}{2\sigma^2/n}\left(\mu-\bar{y}\right)^2\right] \sim N\left(\bar{y},\frac{\sigma^2}{\kappa_n}\right), \kappa_n = n \\
p_J(\sigma^2|\mathbf{y}) &\propto (\sigma^2)^{-(n/2+1)}\exp\left[-\frac{1}{2\sigma^2}\left(n\cdot\frac{\sum_i(y_i-\bar{y})^2}{n}\right)\right] \\
&\propto (\sigma^2)^{-(\nu_n/2+1)}\exp\left[-\frac{1}{2\sigma^2}\left(\nu_n\cdot\sigma_n^2\right)\right] \\
&\sim \text{Inv-}\chi^2(\nu_n,\sigma_n^2), \quad \nu_n = n, \sigma_n^2 = \frac{\sum_i(y_i-\bar{y})^2}{n}
\end{aligned}$$

Hence, this joint density $p_J(\mu,\sigma^2|\mathbf{y})$ can be considered a proper posterior density, which is

$$\text{N-Inv-}\chi^2\left(\bar{y},\frac{\sum_i(y_i-\bar{y})^2}{n^2};n,\frac{\sum_i(y_i-\bar{y})^2}{n}\right)$$

(c)

We can rewrite the prior of (θ,Σ) to

$$p_J(\theta,\Sigma) = C_1|\Sigma|^{-(p+2)/2}$$

in which C_1 is a constant.
Now, we assume that $p_J(\theta,\Sigma)$ is proper, that is

$$\int p_J(\theta,\Sigma)d\theta d\Sigma = \int C_1|\Sigma|^{-(p+2)/2}d\theta d\Sigma = 1$$

By Fubini's Theorem, as well as this post, the marginal distribution of Σ should also be proper (a.s.). However, if we calculate the marginal distribution directly, note that the support of *theta* is $\mathbb{R}^p$

$$\begin{aligned}
p(\Sigma) = \int_{\mathbb{R}^p} p_J(\theta,\Sigma)d\theta &= C_1|\Sigma|^{-(p+2)/2}\int_{\mathbb{R}^p} 1\cdot d\theta \\
&= C_1|\Sigma|^{-(p+2)/2}\cdot\infty \\
&= \infty
\end{aligned}$$

Obviously, the integral above is divergent, which means $p(\Sigma)$ is not proper. By contradiction, $p_J(\theta,\Sigma)$ must be improper. So it cannot actually be a probability density for (θ,Σ).

(d)

Just do it.
Prior

$$p_J(\theta, \Sigma) \propto |\Sigma|^{-(p+2)/2}$$

Likelihood

$$\begin{aligned} p(\mathbf{y}_1, \cdots, \mathbf{y}_n | \theta, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta)\right] \quad y \in \mathbb{R}^p \\ &\propto |\Sigma|^{-n/2} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} S_0\right)\right], \quad S_0 = \sum_{i=1}^{n} (y_i - \theta)(y_i - \theta)^T \end{aligned}$$

Posterior

$$\begin{aligned} p_J(\theta, \Sigma | \mathbf{y}_1, \cdots, \mathbf{y}_n) &\propto p_J(\theta, \Sigma) p(\mathbf{y}_1, \cdots, \mathbf{y}_n | \theta, \Sigma) \\ &\propto |\Sigma|^{-(n+p+2)/2} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} S_0\right)\right] \end{aligned}$$

Note that

$$\begin{aligned} \operatorname{tr}\left(\Sigma^{-1} S_0\right) &= \sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta) \\ &= \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \theta)^T \Sigma^{-1} (y_i - \bar{y} + \bar{y} - \theta) \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma^{-1} (y_i - \bar{y}) - 2\sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma^{-1}(\bar{y} - \theta) + n(\bar{y} - \theta)^T \Sigma^{-1}(\bar{y} - \theta) \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma^{-1} (y_i - \bar{y}) + n(\theta - \bar{y})^T \Sigma^{-1}(\theta - \bar{y}) \\ &= \operatorname{tr}\left(\Sigma^{-1} S\right) + n(\theta - \bar{y})^T \Sigma^{-1}(\theta - \bar{y}), \quad S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T \end{aligned}$$

therefore

$$p_J(\theta, \Sigma | \mathbf{y}_1, \cdots, \mathbf{y}_n) \propto |\Sigma|^{-((n+p)/2+1)} \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Sigma^{-1} S\right) + n(\mu - \bar{y})^T \Sigma^{-1}(\mu - \bar{y})\right)\right]$$

It's known that this term follows a Normal-Inverse-Wishart distribution, formally (using the notation in the book Bayesian Data Analysis Third edition),

$$p_J(\theta, \Sigma | \mathbf{y}_1, \cdots, \mathbf{y}_n) \sim \text{Normal-Inverse-Wishart}\left(\bar{y}, \frac{S}{n}; n, S\right)$$

And we can get

$$p_J(\theta | \Sigma, \mathbf{y}_1, \cdots, \mathbf{y}_n) \propto |\Sigma/n|^{-1/2} \exp\left[-\frac{1}{2}(\theta - \bar{y})^T \left(\frac{\Sigma}{n}\right)^{-1} (\theta - \bar{y})\right] \sim N\left(\bar{y}, \frac{\Sigma}{n}\right)$$

$$p_J(\Sigma | \mathbf{y}_1, \cdots, \mathbf{y}_n) \propto |\Sigma|^{-((n+p+1)/2)} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} S\right)\right] \sim \text{Inv-Wishart}_n(S^{-1})$$