STCS 6071: Probabilistic Models and Machine Learning HW 1

YUCHEN QIU

October 22, 2025

Question 1

Notation Given the Gaussian Mixture model in question, assume known variance σ_k^2 with fixed prior parameters m_0, $\tau_0^2 > 0$, for each $k \in \{1, \cdots, K\}$ define:

$$I_k = \{i : z_i = k\}, \quad N_k = |I_k|, \quad S_k^{(1)} = \sum_{i \in I_k} x_i, \quad \bar{x} = \frac{S_k^{(1)}}{N_k}$$

Part A: Derivations

(a) Label update.

The complete conditional distribution is given by:

$$\begin{aligned} P(z_i = k \mid x_i, \mu_{1:K}) &= \frac{P(z_{1:N}, x_{1:N}, \mu_{1:K})}{P(z_{-i}, x_{1:N}, \mu_{1:K})} \\ &\propto P(z_{1:N} \mid K)\, P(x_{1:N} \mid z_{1:N}, \mu_{1:K})\, P(\mu_{1:K} \mid m_0, \tau_0^2) \\ &\propto P(z_i = k \mid K)\, P(x_i \mid z_i = k, \mu_k) \\ &= P(x_i \mid z_i, \mu_{z_i})\, P(z_i = k \mid K) \\ &= \frac{1}{K} \times \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x_i - \mu_k)^2\right) \end{aligned}$$

Normalize $P(z_i = k | x_i, \mu_{1:K})$ over $k \in \{1, \cdots, K\}$, the normalized categorical distribution is given by:

$$\boxed{P(z_i = k | x_i, \mu_{1:K}) = \frac{(2\pi\sigma_k^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_k^2}(x_i - \mu_k)^2\right)}{\sum_{k=1}^{K} (2\pi\sigma_k^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_k^2}(x_i - \mu_k)^2\right)}}$$

(b) Mean update.

The complete conditional distribution $P(\mu_k | x_{1:K}, z_{1:K})$ is given by:

$$
\begin{aligned}
P(\mu_k|x_{1:K}, z_{1:K}) &= \frac{P(\mathbf{z}, \mathbf{x}, \mu_{1:K})}{P(\mathbf{z}, \mathbf{x})} \\
&\propto P(\mathbf{z}, \mathbf{x}, \mu_{1:K}) \\
&\propto P(\mathbf{z}|K) \times P(\mathbf{x}|\mathbf{z}, \mu_{1:K}) \times P(\mu_{1:K}|m_0, \tau_0^2) \\
&\propto P(\mu_{z_i}|z_i = k, m_0, \tau_0^2) \times \prod_{i \in I_k} P(x_i|z_i, \mu_{z_i}) \\
&= P(\mu_k|m_0, \tau_0^2) \times \prod_{i \in I_k} P(x_i|z_i, \mu_{z_i}) \\
&= \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left(-\frac{1}{2\tau_0^2}(\mu_k - m_0)^2\right) \times \prod_{i \in I_k} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x_i - \mu_k)^2\right) \\
&= (2\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} \exp\left(-\frac{1}{2\tau_0^2}(\mu_k - m_0)^2 - \sum_{i \in I_k} \frac{1}{2\sigma_k^2}(x_i - \mu_k)^2\right) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}(\mu_k^2 - 2\mu_k m_0 + m_0^2) - \frac{1}{2\sigma_k^2}(\sum_{i \in I_k} x_i^2 - 2\sum_{i \in I_k} x_i \mu_k + N_k \mu_k^2)\right) \\
&= \exp\left(-(\frac{1}{2\tau_0^2} + \frac{N_k}{2\sigma_k^2})\mu_k^2 + (\frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2})\mu_k - (\frac{m_0^2}{2\tau_0^2} + \frac{\sum_{i \in I_k} x_i^2}{2\sigma_k^2})\right) \\
&\propto \exp\left(-\frac{1}{2}(\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2})\left[\mu_k^2 - 2\frac{\frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2}}{\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2}}\mu_k + \left(\frac{\frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2}}{\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2}}\right)^2\right] + const\right) \\
&\propto \exp\left(-\frac{1}{2}(\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2})\left(\mu_k - \frac{\frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2}}{\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2}}\right)^2\right)
\end{aligned}
$$

The complete conditional posterior distribution $P(\mu_k|x_{1:K}, z_{1:K})$ has a Gaussian form $\mu_k \mid x_{1:K}, z_{1:K} \sim \mathcal{N}(m_n, \tau_n^2)$:

$$
\boxed{
\begin{aligned}
m_n &= \frac{\frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2}}{\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2}} = \frac{\frac{m_0\sigma_k^2 + S_k^{(1)}\tau_0^2}{\tau_0^2\sigma_k^2}}{\frac{\sigma_k^2 + \tau_0^2 N_k}{\tau_0^2\sigma_k^2}} = \frac{m_0\sigma_k^2 + S_k^{(1)}\tau_0^2}{\sigma_k^2 + \tau_0^2 N_k} \\
\tau_n^2 &= \frac{1}{\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2}} = \frac{1}{\frac{\sigma_k^2 + N_k\tau_0^2}{\tau_0^2\sigma_k^2}} = \frac{\tau_0^2\sigma_k^2}{\sigma_k^2 + N_k\tau_0^2}
\end{aligned}
}
$$

Part B: Collapsed Labels

(c) Cluster-factorized marginal.

Integrating out $\mu_{1:K}$ from joint distribution to obtain $P(x_{1:N}|z_{1:N}$, we factorization the likelihood and prior, and then split the integral by cluster:

$$
\begin{aligned}
p(x_{1:N} \mid z_{1:N}) &= \int \frac{p(x_{1:N}, z_{1:N}, \mu_{1:K})}{p(z_{1:N})}\, d\mu_{1:K} \\
&\propto \int p(x_{1:N} \mid z_{1:N}, \mu_{1:K}) p(z_{1:N}) p(\mu_{1:K} \mid m_0, \tau_0^2)\, d\mu_{1:K} \\
&\propto \int p(x_{1:N} \mid z_{1:N}, \mu_{1:K})\, p(\mu_{1:K} \mid m_0, \tau_0^2)\, d\mu_{1:K} \\
&= \int \prod_{i=1}^{N} p(x_i \mid z_i, \mu_{1:K})\, p(\mu_{1:K} \mid m_0, \tau_0^2)\, d\mu_{1:K} \\
&= \int \prod_{k=1}^{K} \prod_{i \in I_k} p(x_i \mid \mu_k) \prod_{k=1}^{K} p(\mu_k)\, d\mu_k \\
&= \prod_{k=1}^{K} \int \left[\prod_{i \in I_k} p(x_i \mid \mu_k) \right] p(\mu_k)\, d\mu_k \\
&= \prod_{k=1}^{K} p(X_k), \quad X_k = \{x_i : z_i = k\}
\end{aligned}
$$

$$
\begin{aligned}
p(X_k) &= \int \prod_{i \in I_k} p(x_i \mid \mu_k) p(\mu_k)\, d\mu_k \\
&= \int \left[\prod_{i \in I_k} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x_i - \mu_k)^2 \right) \right] \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left(-\frac{1}{2\tau_0^2}(\mu_k - m_0)^2 \right) d\mu_k \\
&= (2\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} \int \exp\left(-\frac{1}{2\tau_0^2}(\mu_k - m_0)^2 - \sum_{i \in I_k} \frac{1}{2\sigma_k^2}(x_i - \mu_k)^2 \right) d\mu_k \\
&= (2\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} \int \exp\left[-\frac{1}{2} \left(\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2} \right) \mu_k^2 + \left(\frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2} \right) \mu_k - \left(\frac{m_0^2}{2\tau_0^2} + \frac{\sum_{i \in I_k} x_i^2}{2\sigma_k^2} \right) \right] d\mu_k \\
&= (2\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} \int \exp\left(-\frac{1}{2} A\mu_k^2 + B\mu_k - C \right) d\mu_k \\
&= (2\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} exp\left(-\frac{1}{2}(C - \frac{B^2}{A}) \right) \int exp\left(-\frac{1}{2} A(\mu_k - \frac{B}{A})^2 \right) d\mu_k \\
&= (2\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} exp\left(-\frac{1}{2}(C - \frac{B^2}{A}) \right) \sqrt{\frac{2\pi}{A}}
\end{aligned}
$$

$$\text{where} \quad A = \frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2}, \quad B = \frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2}, \quad C = \frac{m_0^2}{\tau_0^2} + \frac{\sum_{i \in I_k} x_i^2}{\sigma_k^2}.$$

Specifically,

$$\begin{aligned}
\exp\left(-\frac{1}{2}\Big(C - \frac{B^2}{A}\Big)\right) &= \exp\left(-\frac{1}{2}\left(\frac{m_0^2}{\tau_0^2} + \frac{\sum_{i \in I_k} x_i^2}{\sigma_k^2} - \frac{\left(\frac{m_0}{\tau_0^2} + \frac{S_k^{(1)}}{\sigma_k^2}\right)^2}{\frac{1}{\tau_0^2} + \frac{N_k}{\sigma_k^2}}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i \in I_k} x_i^2}{\sigma_k^2} + \frac{\frac{m_0^2}{\tau_0^2}(\sigma_k^2 + N_k\tau_0^2) - \left(\frac{m_0^2\sigma_k^2}{\tau_0^2} + 2m_0 S_k^{(1)} + \frac{(S_k^{(1)})^2\tau_0^2}{\sigma_k^2}\right)}{\sigma_k^2 + N_k\tau_0^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i \in I_k} x_i^2}{\sigma_k^2} + \frac{m_0^2 N_k - 2m_0 N_k \bar{x}_k - \frac{N_k^2 \bar{x}_k^2 \tau_0^2}{\sigma_k^2}}{\sigma_k^2 + N_k\tau_0^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i \in I_k} x_i^2}{\sigma_k^2} + \frac{-\frac{(\sigma_k^2 + N_k\tau_0^2)N_k\bar{x}_k^2}{\sigma_k^2} + \frac{\sigma_k^2 N_k \bar{x}_k^2}{\sigma_k^2} - 2m_0 N_k \bar{x}_k + m_0^2 N_k}{\sigma_k^2 + N_k\tau_0^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i \in I_k} x_i^2}{\sigma_k^2} - \frac{N_k\bar{x}_k^2}{\sigma_k^2} + \frac{N_k(\bar{x}_k^2 - 2m_0\bar{x}_k + m_0^2)}{\sigma_k^2 + N_k\tau_0^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i \in I_k} x_i^2 - N_k\bar{x}_k^2}{\sigma_k^2} + \frac{N_k(\bar{x}_k - m_0)^2}{\sigma_k^2 + N_k\tau_0^2}\right)\right).
\end{aligned}$$

Therefore, we have the integral that expresses the factor across clusters:

$$\begin{aligned}
p(x_{1:N}|z_{1:N}) &\propto \prod_{k=1}^{K} p(X_k), \quad X_k = \{x_i : z_i = k\} \\
&= \prod_{k=1}^{K} (2\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} \sqrt{\frac{2\pi\tau_0^2\sigma_k^2}{\sigma_k^2 + N_k\tau_0^2}} exp\left(-\frac{1}{2}\left(\frac{\sum_{i \in I_k} x_i^2 - N_k\bar{x}_k^2}{\sigma_k^2} + \frac{N_k(\bar{x}_k - m_0)^2}{\sigma_k^2 + N_k\tau_0^2}\right)\right) \\
&= \prod_{k=1}^{K} (2\pi\sigma_k^2)^{-\frac{N_k}{2}} \sqrt{\frac{\sigma_k^2}{\sigma_k^2 + N_k\tau_0^2}} exp\left(-\frac{\sum_{i \in I_k} x_i^2 - N_k\bar{x}_k^2}{2\sigma_k^2} - \frac{N_k(\bar{x}_k - m_0)^2}{2(\sigma_k^2 + N_k\tau_0^2)}\right) \\
&= (2\pi\sigma_k^2)^{-\frac{N_k}{2}} \sqrt{\frac{\sigma_k^2}{\sigma_k^2 + N_k\tau_0^2}} exp\left(-\frac{\sum_{i \in I_k} x_i^2 - N_k\bar{x}_k^2}{2\sigma_k^2} - \frac{N_k(\bar{x}_k - m_0)^2}{2(\sigma_k^2 + N_k\tau_0^2)}\right)
\end{aligned}$$

(d) Cluster-factorized marginal.

For each $k \in \{1, \cdots, K\}$ define the leave-one-out versions:

$$N_{k,i} = |I_k \setminus \{i\}|, \quad S_{k,-i}^{(1)} = \sum_{i \in I_k \setminus \{i\}} x_i, \quad \bar{x}_{k,-i} = \frac{S_{k,-i}^{(1)}}{N_{k,-i}}$$

The collapsed conditional distribution of z_i given all other variables is

$$p(z_i = k \mid x_{1:N}, z_{-i}) = \frac{p(z_i, z_{-i}, x_{1:N})}{p(z_{-i}, x_{1:N})}.$$

Expanding the numerator and denominator, we obtain

$$\begin{aligned} p(z_i = k \mid x_{1:N}, z_{-i}) &= \frac{p(z_i, z_{-i}, x_{1:N})}{p(z_{-i}, x_{1:N})} \\ &= \frac{p(x_{1:N} \mid z_{1:N})\, p(z_{1:N})}{p(x_i, x_{-i} \mid z_{-i})\, p(z_{-i})} \\ &= \frac{p(x_{1:N} \mid z_{1:N})\, p(z_{1:N})}{p(x_i \mid x_{-i}, z_{-i})\, p(x_{-i} \mid z_{-i})\, p(z_{-i})} \\ &\propto \frac{p(x_{1:N} \mid z_{1:N})\, p(z_{1:N})}{p(x_{-i} \mid z_{-i})\, p(z_{-i})}. \end{aligned}$$

Since the likelihood factorizes over clusters,

$$p(x_{1:N} \mid z_{1:N}) = \prod_{k=1}^{K} p(X_k), \qquad p(x_{-i} \mid z_{-i}) = p(X_k^{(-i)}) \prod_{j \neq k} p(X_j),$$

and the prior $p(z_{1:N})$ is categorical with equal weights $1/K$, it follows that

$$\begin{aligned} p(z_i = k \mid z_{-i}, x_{1:N}) &= \frac{\prod_{k=1}^{K} p(X_k)}{p(X_k^{(-i)}) \prod_{j \neq k} p(X_j)} \times \frac{p(z_{1:N})}{p(z_{-i})} \\ &= \frac{p(X_k)}{p(X_k^{(-i)})} \times \frac{\prod_{k=1}^{K} \frac{1}{K}}{\prod_{j \neq k} \frac{1}{K}} \\ &= \frac{p(X_k)}{p(X_k^{(-i)})} \times \frac{1}{K} \end{aligned}$$

This shows that the collapsed conditional depends on the ratio of marginal likelihoods with and without observation x_i assigned to cluster k, scaled by the prior probability $1/K$.

Based on the result from (c), we have the marginal likelihood $p(x_{1:N} \mid z_{1:N})$ with cluster-specific sets $X_k = \{x_i : z_i = k\}$.

Specifically, the ratio $\frac{p(X_k)}{p(X_k^{(-i)})}$ can be written as

$$\begin{aligned}
\frac{p(X_k)}{p(X_k^{(-i)})} &= \frac{\int \left[\prod_{j\in I_k} p(x_j \mid \mu_k)\right] p(\mu_k)\, d\mu_k}{\int \left[\prod_{j\in I_k\setminus\{i\}} p(x_j \mid \mu_k)\right] p(\mu_k)\, d\mu_k} \\
&= \frac{\int p(X_k^{(-i)} \mid \mu_k)\, p(x_i \mid \mu_k)\, p(\mu_k)\, d\mu_k}{\int p(X_k^{(-i)} \mid \mu_k)\, p(\mu_k)\, d\mu_k} \\
&= \int p(x_i \mid \mu_k)\, \frac{p(X_k^{(-i)} \mid \mu_k)\, p(\mu_k)}{\int p(X_k^{(-i)} \mid \mu_k)\, p(\mu_k)\, d\mu_k}\, d\mu_k \\
&= \int p(x_i \mid \mu_k)\, p(\mu_k \mid X_k^{(-i)})\, d\mu_k \\
&= p(x_i \mid X_k^{(-i)}),
\end{aligned}$$

which is exactly the posterior predictive distribution of x_i given the other data currently assigned to cluster k.

The likelihood for a new data point is $x_i \mid \mu_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$; the posterior predicted distribution is still Gaussian:

$$x_i | X^{(-i)} \sim \mathcal{N}(m_{n,-i}, \sigma_k^2 + \tau_{n,-i}^2)$$

Use $X_k^{(-i)}$ to compute posterior distribution:

$$\begin{aligned}
\mu_k \mid x_{-i}, z_{-i} &\sim \mathcal{N}(m_{n_{k,-i}}, \tau_{n_{k,-i}}^2) \\
where \quad m_{n_{k,-i}} &= \frac{m_0\sigma_k^2 + S_{k,-i}^{(1)}\tau_0^2}{\sigma_k^2 + \tau_0^2 N_{k,-i}}, \quad \tau_{n_{k,-i}}^2 = \frac{\tau_0^2\sigma_k^2}{\sigma_k^2 + N_{k,-i}\tau_0^2}
\end{aligned}$$

Then,

$$\begin{aligned}
p(z_i = k | z_{-i}, x_{1:N}) &\propto p(x_i \mid X_k^{(-i)}) \times \frac{1}{K} \\
&= \frac{1}{K\sqrt{2\pi(\sigma_k^2 + \tau_{n_{k,-i}}^2)}} exp\left(-\frac{(x_i - m_{n_{k,-i}})^2}{2(\sigma_k^2 + \tau_{n_{k,-i}}^2)}\right)
\end{aligned}$$

Normalize the single site collapsed conditional over $k \in \{1, \cdots, K\}$:

$$p(z_i = k | z_{-i}, x_{1:N}) = \frac{\frac{1}{K} p(x_i \mid X_k^{(-i)})}{\sum_{j=1}^K \frac{1}{K} p(x_i \mid X_j^{(-i)})}$$

When update z_i in a collapsed Gibbs sampler,

- when assigning i to a nonempty cluster ($N_{k,-i} \geq 1$), we compute the posterior predictive $p(x_i | X_k^{(-i)})$ directly, $p(x_i | X_k^{(-i)}) = \mathcal{N}(m_{n_{k,-i}}, \sigma_k^2 + \tau_{n_{k,-i}}^2)$

- when assigning i to an empty cluster ($N_{k,-i} = 0$), there is no data informing μ_k, the posterior for μ_k is just the piror $\mathcal{N}(m_0, \tau_0^2)$, so the predictive is the prior predictive $p(x_i|X_k^{(-i)}) = \mathcal{N}(m_0, \sigma_k^2 + \tau_0^2)$

The normalized collapsed conditional distribution for z_i is:

$$p(z_i = k \mid z_{-i}, x_{1:N}) = \frac{p(x_i \mid X_k^{(-i)})}{\sum_{j=1}^{K} p(x_i \mid X_j^{(-i)})},$$

$$p(x_i \mid X_k^{(-i)}) = \begin{cases} \mathcal{N}(m_{n_{k,-i}},\, \sigma_k^2 + \tau^2_{n_{k,-i}}) & N_{k,-i} \geq 1 \\ \mathcal{N}(m_0,\, \sigma_k^2 + \tau_0^2) & N_{k,-i} = 0 \end{cases}$$

$$m_{n_{k,-i}} = \frac{m_0 \sigma_k^2 + S^{(1)}_{k,-i} \tau_0^2}{\sigma_k^2 + \tau_0^2 N_{k,-i}}, \quad \tau^2_{n_{k,-i}} = \frac{\tau_0^2 \sigma_k^2}{\sigma_k^2 + N_{k,-i} \tau_0^2}$$

Part D: Graphical Depiction

(e) Depict Graphical Models

The plate notation for the graphical model is shown in Figure 1.

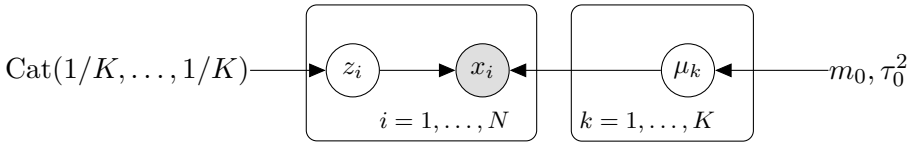


Figure 1: Graphical model for the collapsed Gibbs sampler.

Question 2

Notation. For each k, let $I_k = \{i : z_i = k\}$, $N_k = |I_k|$, and define the sufficient statistics

$$S_k^{(1)} = \sum_{i \in I_k} x_i, \qquad S_k^{(2)} = \sum_{i \in I_k} x_i^2, \qquad \bar{x}_k = \frac{S_k^{(1)}}{N_k} \quad \text{(defined if } N_k \geq 1\text{)}.$$

For single–site label updates, the leave–one–out summaries $N_{k,-i}, S_{k,-i}^{(1)}, S_{k,-i}^{(2)}$, computed on $I_k \setminus \{i\}$, and $\bar{x}_{k,-i} = \frac{S_{k,-i}^{(1)}}{N_{k,-i}}$ (defined if $N_{k,-i} \geq 1$).

Part A: Blocked Gibbs (derive complete conditionals; normalized)

(a) Weights

For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k > 0$, the probability density for the vector $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_k)$:

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1},$$

where $B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$ is a normalizing constant.

The posterior distribution of the mixing proportions $\boldsymbol{\pi}$ given the latent assignments $z_{1:N}$ is:

$$\begin{aligned} p(\boldsymbol{\pi}|z_{1:N}) &\propto p(\boldsymbol{\pi}, z_{1:N}) = p(z_{1:N} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \\ &= \prod_{k=1}^{K} \prod_{i \in I_k} \pi_k \times \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \\ &\propto \prod_{k=1}^{K} \pi_k^{N_k + \alpha_k - 1} \end{aligned}$$

Therefore the posterior is also a Dirichlet distribution with posterior parameters:

$$\boxed{\pi \mid z_{1:N} \sim \mathrm{Dir}(_1 + N_1,\ \ldots,\ _K + N_K), \quad N_k = |I_k|}$$

(b) Means

$$\begin{aligned} p(\mu_k|\lambda_k, x_{1:N}, z_{1:N}) &\propto p(\mu_{1:K}, \lambda_{1:K}, x_{1:N}, z_{1:N}, \pi) \\ &\propto p(x_i \mid z_i = k, \mu_k, \lambda_k) p(\mu_k \mid \lambda_k) \\ &= \prod_{i \in I_k} p(x_i \mid \mu_k, \lambda_k) \times p(\mu_k \mid m_0, \kappa_0, \lambda_k) \\ &\propto \exp\left[-\frac{\sum_{i \in I_k} (x_i - \mu_k)^2}{2\lambda_k^{-1}} - \frac{(\mu_k - m_0)^2}{2(\kappa_0 \lambda_k)^{-1}} \right] \\ &\propto \exp\left[-\left(\frac{\kappa_0 \lambda_k}{2} + \frac{\lambda_k N_k}{2} \right) \mu_k^2 + \left(m_0 \kappa_0 \lambda_k + \lambda_k S_k^{(1)} \right) \mu_k - const \right] \\ &\propto \exp\left[-\frac{1}{2[\lambda_k(\kappa_0 + N_k)]^{-1}} \left(\mu_k - \frac{m_0 \kappa_0 + S_k^{(1)}}{\kappa_0 + N_k} \right)^2 \right] \end{aligned}$$

The complete conditional posterior distribution of μ_k is also Gaussian $\boxed{\mu_k \mid \lambda_k, x_{1:N}, z_{1:N} \sim \mathcal{N}(m_n, (\kappa_n \lambda_k)^{-1})}$,

where $\boxed{m_n = \frac{m_0 \kappa_0 + S_k^{(1)}}{\kappa_0 + N_k}, \quad \kappa_n = \kappa_0 + N_k}$.

(c) Precisions

$$\begin{aligned}
p(\lambda_k \mid \mu_k, x_{1:N}, z_{1:N}) &\propto p(\mu_{1:K}, \lambda_{1:K}, x_{1:N}, z_{1:N}, \pi) \\
&\propto p(x_i \mid z_i = k, \mu_k, \lambda_k) p(\mu_k \mid \lambda_k) p(\lambda_k) \\
&= \prod_{i \in I_k} [2\pi\lambda_k^{-1}]^{-\frac{1}{2}} \exp\left[-\frac{(x_i - \mu_k)^2}{2\lambda_k^{-1}}\right] \times [2\pi(\kappa_0\lambda_k)^{-1}]^{\frac{1}{2}} \exp\left[-\frac{(\mu_k - m_0)^2}{2(\kappa_0\lambda_k)^{-1}}\right] \\
&\times \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda_k^{a_0 - 1} \exp(-b_0 \lambda_k) \\
&\propto \lambda_k^{\frac{N_k}{2} + \frac{1}{2} + a_0 - 1} \exp\left[\left(-\frac{S_k^{(2)} - 2S_k^{(1)}\mu_k + N_k\mu_k^2}{2} - \frac{\kappa_0(\mu_k - m_0)^2}{2} - b_0\right)\lambda_k\right]
\end{aligned}$$

The complete conditional posterior distribution of λ_k is also Gamma distribution:

$$\boxed{\lambda_k \mid \mu_k, x_{1:N}, z_{1:N} \sim \text{Gamma}(\frac{N_k + 1}{2} + a_0, \frac{S_k^{(2)} - 2S_k^{(1)}\mu_k + N_k\mu_k^2 + \kappa_0(\mu_k - m_0)^2}{2} + b_0)}$$

(d) Labels

$$\begin{aligned}
p(z_i = k \mid x_i, \mu_{1:K}, \lambda_{1:K}, \pi) &\propto p(\mu_{1:K}, \lambda_{1:K}, x_{1:N}, z_{1:N}, \pi) \\
&\propto p(x_i \mid z_i = k, \mu_k, \lambda_k) p(z_i = k \mid \pi) \\
&= (2\pi\lambda_k^{-1})^{-\frac{1}{2}} \exp\left[-\frac{(x_i - \mu_k)^2}{2\lambda_k^{-1}}\right] \times \pi_k
\end{aligned}$$

The normalized single–site categorical distribution over $k = 1, \cdots, K$ is given by:

$$\boxed{p(z_i = k \mid x_i, \mu_{1:K}, \lambda_{1:K}, \pi) = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \lambda_j^{-1})}{\sum_{j=1}^K \pi_j \mathcal{N}(x_i | \mu_j, \lambda_j^{-1})}}$$

Part B: Component–collapsed labels

Notation. When updating z_i, define the current contents of cluster k excluding i by $D_{k,-i} = \{\, x_j : z_j = k,\ j \neq i \,\}$, with corresponding summaries $N_{k,-i}, S_{k,-i}^{(1)}, S_{k,-i}^{(2)}$, and $\bar{x}_{k,-i} = \frac{S_{k,-i}^{(1)}}{N_{k,-i}}$ when $N_{k,-i} \geq 1$.

(e) Posterior for component k given $D_{k,-i}$(up to proportionality).

$$
\begin{aligned}
p(\mu_k, \lambda_k \mid D_{k,-i}) &= \frac{p(\mu_k, \lambda_k, D_{k,-i})}{p(D_{k,-i})} \\
&\propto p(D_{k,-i} \mid \mu_k, \lambda_k) p(\mu_k \mid \lambda_k) p(\lambda_k) \\
&= \prod_{j \in I_k \setminus \{i\}} (2\pi\lambda_k^{-1})^{-\frac{1}{2}} \exp\left[-\frac{(x_i - \mu_k)^2}{2\lambda_k^{-1}} \right] \times (2\pi(\kappa_0\lambda_k)^{-1})^{-\frac{1}{2}} \exp\left[-\frac{(\mu_k - m_0)^2}{2(\kappa_0\lambda_k)^{-1}} \right] \\
&\times \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda_k^{a_0 - 1} \exp(-b_0\lambda_k) \\
&\propto \lambda_k^{\frac{N_{k,-i}+1}{2} + a_0 - 1} \exp\left[\left(-\frac{S^{(2)}_{k,-i} - 2S^{(1)}_{k,-i}\mu_k + N_{k,-i}\mu_k^2}{2} - \frac{\kappa_0(\mu_k - m_0)^2}{2} - b_0 \right) \lambda_k \right] \\
&= \lambda_k^{\frac{N_{k,-i}+1}{2} + a_0 - 1} \exp\left[-\left(\frac{(N_{k,-i} + \kappa_0)\mu_k^2 - 2(S^{(1)}_{k,-i} + \kappa_0 m_0)\mu_k + S^{(2)}_{k,-i} + \kappa_0 m_0^2}{2} + b_0 \right) \lambda_k \right] \\
&= (\lambda_k^{-1})^{-\frac{1}{2}} \exp\left[-\frac{(\mu_k - \frac{S^{(1)}_{k,-i} + \kappa_0 m_0}{N_{k,-i} + \kappa_0})^2}{2[(N_{k,-i} + \kappa_0)\lambda_k]^{-1}} \right] \\
&\times \lambda_k^{\frac{N_{k,-i}}{2} + a_0 - 1} \exp\left[-\left(-\frac{(S^{(1)}_{k,-i} + \kappa_0 m_0)^2}{2(N_{k,-i} + \kappa_0)} + \frac{S^{(2)}_{k,-i} + \kappa_0 m_0^2}{2} + b_0 \right) \lambda_k \right]
\end{aligned}
$$

The joint posterior stays in the Normal–Gamma family, the updated constants for the posterior are:

$$
\boxed{
\begin{aligned}
m_{k,i} &= \frac{S^{(1)}_{k,-i} + \kappa_0 m_0}{N_{k,-i} + \kappa_0}, \quad \kappa_{k,i} = N_{k,-i} + \kappa_0 \\
a_{k,-i} &= \frac{N_{k,-i}}{2} + a_0, \quad b_{k,-i} = \frac{1}{2}[S^{(2)}_{k,-i} + \kappa_0 m_0^2 - \frac{(S^{(1)}_{k,-i} + \kappa_0 m_0)^2}{N_{k,-i} + \kappa_0}] + b_0
\end{aligned}
}
$$

Case $N_{k,-i} = 0$. There are no observations in cluster k after removing i. Hence the updated constants are:

$$
\boxed{m_{k,-i} = m_0, \quad \kappa_{k,-i} = \kappa_0, \quad a_{k,-i} = a_0, \quad b_{k,-i} = b_0,}
$$

so the posterior over (μ_k, λ_k) is just the prior, and the predictive for assigning x_i to an empty cluster is the prior predictive.

(f) Posterior predictive under component k.

The posterior predictive distribution:

$$
\begin{aligned}
p_{pred}(x_i \mid D_{k,-i}) &= \iint \mathcal{N}(x_i \mid \mu_k, \lambda_k^{-1}) p(\mu_k, \lambda_k \mid d_{k,-i}) d\mu_k d\lambda_k \\
&= \int p(\lambda_k \mid D_{k,-i}) \left[\int \mathcal{N}(x_i \mid \mu_k, \lambda_k^{-1}) \mathcal{N}(\mu_k \mid m_{k,-i}, (\kappa_{k,-i}\lambda_k)^{-1}) d\mu_k \right] d\lambda_k \\
&= \int p(\lambda_k \mid D_{k,-i}) \times p(x_i \mid \lambda_k, D_{k,-i}) d\lambda_k
\end{aligned}
$$

First integrate out μ_k:

$$
\begin{aligned}
p(x_i \mid \lambda_k, D_{k,-i}) &= \int \mathcal{N}(x_i \mid \mu_k, \lambda_k^{-1}) \mathcal{N}(\mu_k \mid m_{k,-i}, (\kappa_{k,-i}\lambda_k)^{-1}) d\mu_k \\
&= \int \sqrt{\frac{\lambda_k}{2\pi}} \exp\left(-\frac{\lambda_k}{2\pi}(x_i - \mu_k)^2\right) \times \sqrt{\frac{\kappa_{k,-i}\lambda_k}{2\pi}} \exp\left(-\frac{\kappa_{k,-i}\lambda_k}{2\pi}(\mu_k - m_{k,-i})^2\right) d\mu_k \\
&= \int \frac{\lambda_k \sqrt{\kappa_{k,-i}}}{2\pi} \exp\left(-\frac{\lambda_k}{2}\left[(x_i - \mu_k)^2 + \kappa_{k,-i}(\mu_k - m_{k,-i})^2\right]\right) \\
&= \frac{\lambda_k \sqrt{\kappa_{k,-i}}}{2\pi} \exp\left\{-\frac{\lambda_k}{2}\left[x_i^2 + \kappa_{k,-i} m_{k,-i}^2 - \frac{(x_i + \kappa_{k,-i} m_{k,-i})^2}{1 + \kappa_{k,-i}}\right]\right\} \\
&\quad \times \int \exp\left[-\frac{\lambda_k}{2}(1 + \kappa_{k,-i})\left(\mu_k - \frac{x_i + \kappa_{k,-i} m_{k,-i}}{1 + \kappa_{k,-i}}\right)^2\right] d\mu_k \\
&= \frac{\lambda_k \sqrt{\kappa_{k,-i}}}{2\pi} \sqrt{\frac{2\pi}{\lambda_k(1 + \kappa_{k,-i})}} \exp\left\{-\frac{\lambda_k}{2}\left[x_i^2 + \kappa_{k,-i} m_{k,-i}^2 - \frac{(x_i + \kappa_{k,-i} m_{k,-i})^2}{1 + \kappa_{k,-i}}\right]\right\} \\
&= \sqrt{\frac{\lambda_k}{2\pi\,(1 + 1/\kappa_{k,-i})}} \exp\left[-\frac{\lambda_k}{2(1 + 1/\kappa_{k,-i})}(x_i - m)^2\right]
\end{aligned}
$$

Therefore, $p(x_i \mid \lambda_k, D_{k,-i})$ is still normal $x_i \mid \lambda_k, D_{k,-i} \sim \mathcal{N}(m_{k,-i}, \lambda_k^{-1}(1 + \frac{1}{\kappa_{k,-i}})$.

Then integrate out λ_k:

$$
\begin{aligned}
p(x_i \mid D_{k,-i}) &= \int p(\lambda_k \mid D_{k,-i}) \times p(x_i \mid \lambda_k, D_{k,-i}) d\lambda_k \\
&= \int \frac{b_{k,-i}^{a_{k,-i}}}{\Gamma(a_{k,-i})} \lambda_k^{a_{k,-i}-1} \exp\left(-b_{k,-i}\lambda_k\right) \times \sqrt{\frac{\lambda_k}{2\pi\,(1 + 1/\kappa_{k,-i})}} \exp\left[-\frac{\lambda_k}{2(1 + 1/\kappa_{k,-i})}(x_i - m)^2\right] d\lambda_k \\
&= \frac{b_{k,-i}^{a_{k,-i}}}{\Gamma(a_{k,-i})} \frac{1}{\sqrt{2\pi\,(1 + 1/\kappa_{k,-i})}} \int \lambda_k^{a_{k,-i} + \frac{1}{2} - 1} \exp\left[\lambda_k\left(b_{k,-i} + \frac{(x_i - m_{k,-i})^2}{2(1 + 1/\kappa_{k,-i})}\right)\right] d\lambda_k \\
&= \frac{b_{k,-i}^{a_{k,-i}}}{\Gamma(a_{k,-i})} \frac{1}{\sqrt{2\pi\,(1 + 1/\kappa_{k,-i})}} \frac{\Gamma(a_{k,-i} + \frac{1}{2})}{(b_{k,-i} + \frac{(x_i - m_{k,-i})^2}{2(1 + 1/\kappa_{k,-i})})^{a_{k,-i} + \frac{1}{2}}} \\
&= \frac{\Gamma(a_{k,-i} + \frac{1}{2})}{\Gamma(a_{k,-i})} \frac{1}{\sqrt{\pi 2 b_{k,-i}(1 + 1/\kappa_{k,-i})}} \left[1 + \frac{(x_i - m_{k,-i})^2}{2 b_{k,-i}(1 + 1/\kappa_{k,-i})}\right]^{-\frac{2a_{k,-i}+1}{2}}
\end{aligned}
$$

Let $v_{k,-i} = 2a_{k,-i}$, $s_{k,-i}^2 = \frac{b_{k,-i}}{a_{k,-i}}(1 + \frac{1}{\kappa_{k,-i}})$, the posterior predictive distribution is Student-t distribution $\boxed{x_i \mid D_{k,-i} \sim \text{Student-t}_{2a_{k,-i}}\left(m_{k,-i}, \frac{b_{k,-i}}{a_{k,-i}}(1 + 1/\kappa_{k,-i})\right)}$, where

$$
m_{k,i} = \frac{S_{k,-i}^{(1)} + \kappa_0 m_0}{N_{k,-i} + \kappa_0}, \quad \kappa_{k,i} = N_{k,-i} + \kappa_0, \quad a_{k,-i} = \frac{N_{k,-i}}{2} + a_0, \quad b_{k,-i} = \frac{1}{2}[S_{k,-i}^{(2)} + \kappa_0 m_0^2 - \frac{(S_{k,-i}^{(1)} + \kappa_0 m_0)^2}{N_{k,-i} + \kappa_0}] + b_0
$$

Case $N_{k,-i} = 0$. If the cluster is empty after removing i, the predictive reduces to the prior predictive:

$$
\boxed{x_i \mid N_{k,-i} = 0 \sim \text{Student-t}_{2a_0}\left(m_0, \frac{b_0}{a_0}(1 + 1/\kappa_0)\right)}
$$

(g) Component–collapsed label conditional.

The complete conditional for the single-site update is given by:

$$
\begin{aligned}
p(z_i = k \mid x_{1:N}, z_{-i}, \pi) &\propto \frac{p(z_{1:N}, x_{1:N}, \pi)}{p(z_{-i}, D_{k,-i}, \pi)} \\
&= \frac{p(x_{1:N} \mid z_{1:N}, \pi)\, p(z_{1:N} \mid \pi)}{p(D_{k,-i} \mid z_{-i}, \pi)\, p(z_{-i} \mid \pi)} \\
&= p(x_i \mid D_{k,-i}) \times p(z_i = k \mid \pi) \\
&= \pi_k \times \underbrace{p(x_i \mid D_{k,-i})}_{\text{Student-}t\text{ density}}
\end{aligned}
$$

$$
\text{where} \quad p(x_i \mid D_{k,-i}) = \begin{cases} t_{2a_{k,-i}}(x_i; m_{k,-i}, \frac{b_{k,-i}}{a_{k,-i}}(1 + 1/\kappa_{k,-i})), & N_{k,-i} \geq 1 \\ t_{2a_0}\left(m_0, \frac{b_0}{a_0}(1 + 1/\kappa_0)\right) & N_{k,-i} = 0 \end{cases}
$$

$$
m_{k,i} = \frac{S^{(1)}_{k,-i} + \kappa_0 m_0}{N_{k,-i} + \kappa_0}, \quad \kappa_{k,i} = N_{k,-i} + \kappa_0, \quad a_{k,-i} = \frac{N_{k,-i}}{2} + a_0, \quad b_{k,-i} = \frac{1}{2}[S^{(2)}_{k,-i} + \kappa_0 m_0^2 - \frac{(S^{(1)}_{k,-i} + \kappa_0 m_0)^2}{N_{k,-i} + \kappa_0}] + b_0
$$

Part C: What changed relative to Question 1

In Q1, labels ignore mixture weights and use fixed, equal priors, so the blocked label conditional is proportional to the Gaussian likelihood alone; in Q2, the prior weight π_k tilts the probabilities toward clusters that already have more points, which means large clusters grow more easily.

In Q1, when updating μ_k, the posterior variance depends only on N_k. In Q2, the prevision λ_k is learned, so the posterior variance of μ_k depends on both N_k and uncertainty in λ_k. In Q1, the label updates depend on likelihood with fixed width; in Q2, learned λ_k allows each cluster to adapt its spread to the data, which tightens the posterior when the cluster is well-formed.

In Q2, integrating out μ_k, λ_k yields a Student-t predictive distribution with heavier tails. Compared to Normal predictive in Q1, this makes small or early-stage clusters less sensitive to outliers, allowing more robust assignments and typically improving label mixing when $N_{k,-i}$ is small.

Part D: A graphical depiction

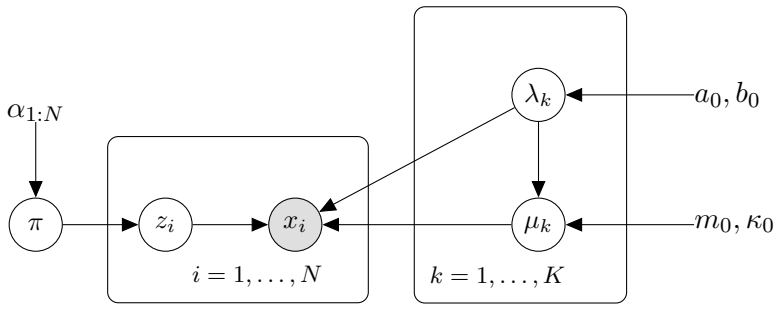


Figure 2: Plate notation for a Gaussian mixture modelwith unknown variance.

Question 3

Implementation — Collapsed vs. Blocked Mixture Inference

Setup: S2 (Learned weights & variances)

Model.

Let $x_{1:N} \in \mathbb{R}$ and fix the number of components $K \geq 2$. We assume the following hierarchical mixture model:

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K), \tag{1}$$

$$z_i \mid \boldsymbol{\pi} \sim \mathrm{Cat}(\pi_1, \ldots, \pi_K), \quad i = 1, \ldots, N, \tag{2}$$

$$x_i \mid (z_i = k, \mu_k, \lambda_k) \sim \mathcal{N}(\mu_k, \lambda_k^{-1}), \tag{3}$$

$$\lambda_k \sim \mathrm{Gamma}(a_0, b_0), \tag{4}$$

$$\mu_k \mid \lambda_k \sim \mathcal{N}\big(m_0, (\kappa_0 \lambda_k)^{-1}\big), \quad k = 1, \ldots, K, \tag{5}$$

with fixed prior parameters $\alpha_k > 0$, $a_0, b_0, \kappa_0 > 0$, and $m_0 \in \mathbb{R}$. Use the shape–rate parameterization for the Gamma distribution $p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$.

Data Generating Process.

Synthetic data is generated according to the model:

1. Fix the number of components K and sample mixture weights
$$\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K).$$
2. For each component $k = 1, \ldots, K$, draw
$$\lambda_k \sim \mathrm{Gamma}(a_0, b_0), \qquad \mu_k \mid \lambda_k \sim \mathcal{N}\big(m_0, (\kappa_0 \lambda_k)^{-1}\big).$$
3. For each observation $i = 1, \ldots, N$:
   (a) Assign a cluster label $z_i \sim \mathrm{Cat}(\boldsymbol{\pi})$.
   (b) Generate $x_i \sim \mathcal{N}(\mu_{z_i}, \lambda_{z_i}^{-1})$.

Hyperparameter Setting.

Set hyperparameter for data generation using the following configuration:

$$\text{Cluster number} : K = 5, \qquad \text{Data sample size} : N = 1000,$$

$$\text{Dirichlet prior for mixture weights} : \boldsymbol{\alpha} = (0.5,\ 1,\ 2,\ 3,\ 5),$$

$$\text{Normal prior hyperparameters} : m_0 = 1, \qquad \kappa_0 = 0.25,$$

$$\text{Gamma(shape–rate) prior for precisions } \lambda_k : a_0 = 10, \qquad b_0 = 2,$$

When run Gibbs samplers, set the same cluster number $K = 5$ and Dirichlet prior as for data generating process; initialize hyperparameter of Normal prior as $m_0 = 0$, $\kappa_0 = 0.5$, and hyperparameter of Gamma prior as $a_0 = 3$, $b_0 = 3$; set burn-in period $S_{\text{burn-in}} = 1000$, keep chain length $S_{\text{chain}} = 10000$.

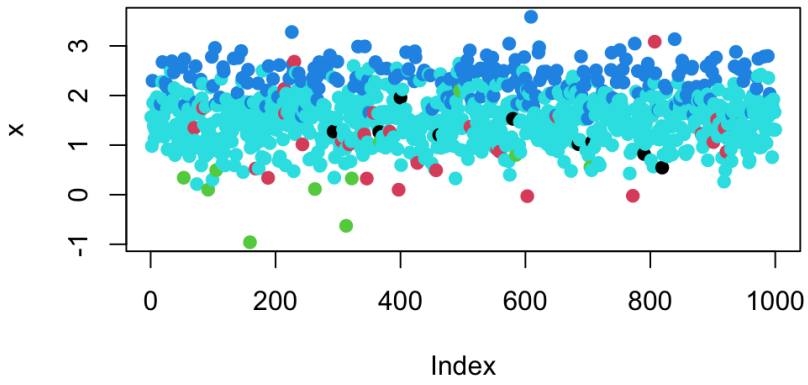
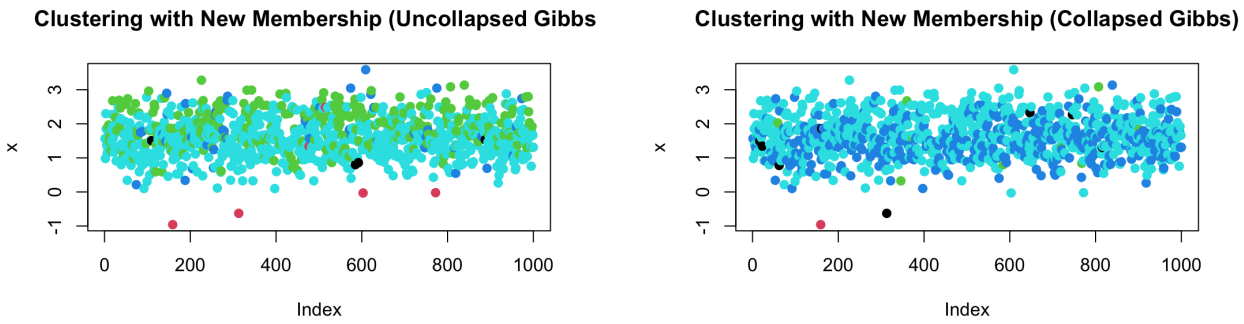


Figure 3: True membership of synthetic data

Visualization: fitted clustering

The fitted values using blocked Gibbs sampler are as follows:



(a) Blocked Gibbs Sampler

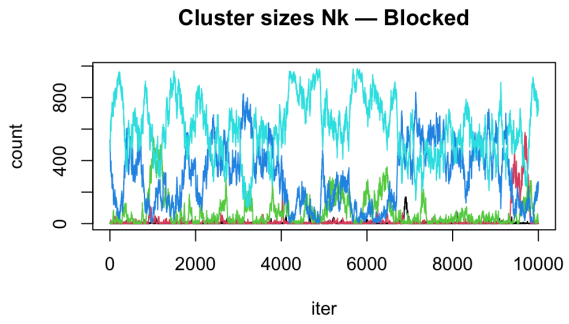
(b) Collapsed Gibbs Sampler

Figure 4: Fitted Clustering Results

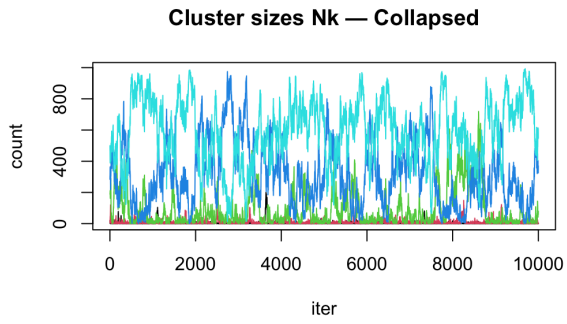
Diagnostic of cluster sizes N_k

Trace Plotting for Two Gibbs Sampler

The visualization of the trace of cluster size N_k of two chains are as follows:



(a) Caption for the first figure

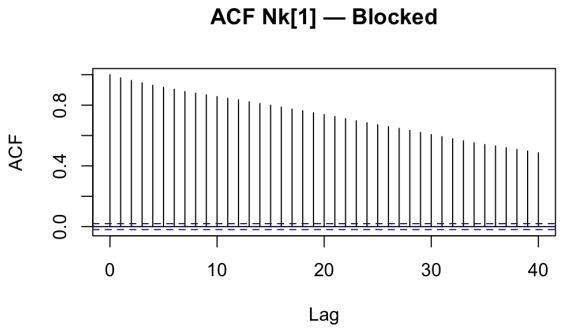


(b) Caption for the second figure

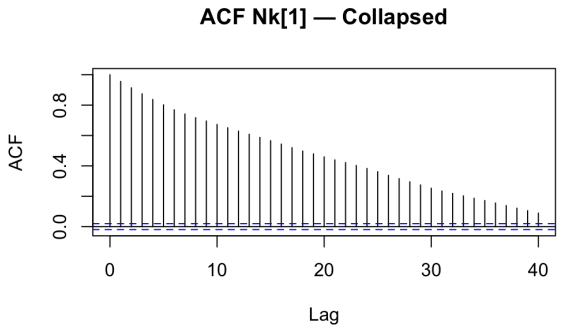
Figure 5: Overall caption describing both subfigures.

ACF Plotting for Cluster 1

Select cluster 1 to visualize the autocorrelation function (ACF) for both sampling methods, illustrating how the value of the chain at iteration 4 correlates with its past values at lag ℓ.



(a) Caption for the first figure



(b) Caption for the second figure

Figure 6: Overall caption describing both subfigures.

Discussion

From trace plotting, we found that:

Blocked Gibbs Sampler

- The traces fluctuate a lot, and the chain moves slowly between different cluster configurations, which is visible as long plateaus and slower transitions.
- This reflects strong autocorrelation and slower mixing, the sampler tends to "stick" in a few configurations before jumping to another.

Collapsed Gibbs Sampler

- The traces still fluctuate, but transitions between cluster configurations are faster and more frequent. This indicates weaker autocorrelation and faster mixing.
- No cluster remains dominant for long stretches — the sampler explores different partitions more readily.

From ACF plotting, we found that:

Collapsed Gibbs Sampler

- The ACF drops quickly, indicating faster decorrelation between successive samples. This suggests that the chain mixes more efficiently, explores the posterior more quickly, and achieves a higher effective sample size.

Blocked Gibbs Sampler

- The ACF decays more slowly, revealing stronger autocorrelation in the cluster size trajectory. This implies that successive samples are more dependent, leading to slower mixing, lower effective sample size, and less efficient exploration of the posterior.

Question 3.1: GMM Implementation & Mini-Report

Modeling Track Declaration: Track C (non-Gaussian mixture model)

Introduction

We want to study the latent preference of influencers on 500 popular brands using an influencer-brand social media dataset. Assume each influencer's observed interactions across brands as a draw from a cluster-specific brand preference distribution.

Data

The data is the interaction records between 1928 influencers and 500 top mentioned brands from instagram influencer dataset from Kim et al. (2020). The data matrix $X \in \mathbb{N}^{N\times D}$ includes $N = 1928$ influencers and $D = 500$ brands, x_{ij} is the number of times influencer i mentioned brand i in the historical content.

Model: Dirichlet–Multinomial Mixture

Apply Multinomial-Dirichlet Mixture model in capturing latent brand preference across influencers.

Assume:

1. Each influencer belongs to some latent cluster ($z\_i \in \{1, \dots, K\}$).

2. Each cluster k is a mixture of brands, which represents its own brand preference distribution:

$$\boldsymbol{\phi}_k = (\phi_{k1}, \cdots, \phi_{kD}), \quad \sum_{j=1}^{D} \phi_{kj} = 1.$$

3. Given $z_i = k$, the interactions for influencer i are generated as $\mathbf{x}_i \sim \text{Multinomial}(n_i, \boldsymbol{\phi}_k)$, $\mathbf{x}_i = (x_{i1}, , \cdots, x_{iD})$.

In each cluster, influencers have a specific preference in interacting with some specific category of brands. For example, cluster 1 might heavily interact with fashion brands, cluster 2 might prefer gaming brands, etc.

Setup

Data. For each influencer $i = 1, \dots, N$, let $\mathbf{x}_i = (x_{i1}, \dots, x_{iD})^\top \in \mathbb{N}_0^D$ be interaction counts over D brands, and let $n_i = \sum_{j=1}^{D} x_{ij}$.

Latent labels and parameters. Fix $K \geq 2$ clusters. Let $z_i \in \{1, \dots, K\}$ be the latent segment for influencer i. Each cluster k has a brand-preference vector $\boldsymbol{\phi}_k = (\phi_{k1}, \dots, \phi_{kD})^\top$ on the D-simplex, and mixture weights $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^\top$.

Priors and likelihood

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K), \tag{6}$$

$$\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta_1, \ldots, \beta_D), \quad k = 1, \ldots, K, \tag{7}$$

$$z_i \mid \boldsymbol{\pi} \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad i = 1, \ldots, N, \tag{8}$$

$$\mathbf{x}_i \mid z_i = k, \boldsymbol{\phi}_k \sim \mathrm{Multinomial}(n_i, \boldsymbol{\phi}_k). \tag{9}$$

with fixed α, β and K.

Joint distribution.

$$p(\mathbf{x}, \mathbf{z}, \{\boldsymbol{\phi}_k\}, \boldsymbol{\pi}) = p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{\phi}_k) \prod_{i=1}^{N} \left[p(z_i \mid \boldsymbol{\pi})\, p(\mathbf{x}_i \mid z_i, \{\boldsymbol{\phi}_k\}) \right].$$

Posterior Inference

Let $N_k = \#\{i : z_i = k\}$ and $C_{kj} = \sum_{i:z_i=k} x_{ij}$ (cluster–brand counts).

Blocked Gibbs Updates:

$$\boldsymbol{\pi} \mid \mathbf{z} \sim \mathrm{Dir}\big(\alpha_1 + N_1, \ldots, \alpha_K + N_K\big), \tag{10}$$

$$\boldsymbol{\phi}_k \mid \{\mathbf{x}_i : z_i = k\} \sim \mathrm{Dir}\big(\beta_1 + C_{k1}, \ldots, \beta_D + C_{kD}\big), \tag{11}$$

$$p(z_i = k \mid \cdot) \propto \pi_k \cdot \mathrm{Mult}\big(\mathbf{x}_i \mid n_i, \boldsymbol{\phi}_k\big) \propto \pi_k \prod_{j=1}^{D} \phi_{kj}^{x_{ij}} \tag{12}$$

Empty cluster: if $N_{k,-i} = 0$, use the prior predictive by setting $C_{kj,-i} = 0$.

Graphical depiction

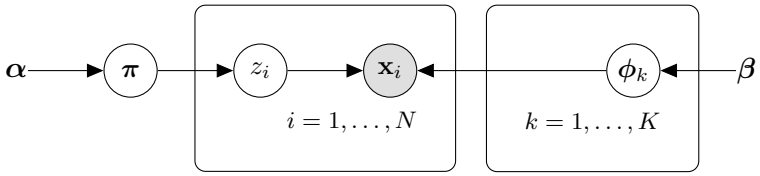


Figure 7: Plate notation for a Dirichlet–Multinomial mixture model.

Results

Set symmetric priors $\alpha_k = \alpha_0 = 1$, $\beta_k = \beta_0 = 0.5$, $K = 20$ and burn-in perior =1000 runs. Run $S = 10000$. The posterior mean of mixture weights are as follows:

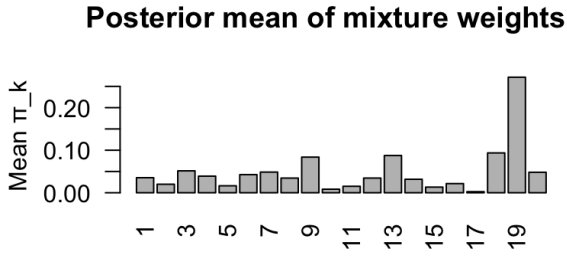


Figure 8: Posterior Mean of Brand Cluster Mixture Weights

I picked cluster1, cluster9, cluster19's top 10 brand accounts:

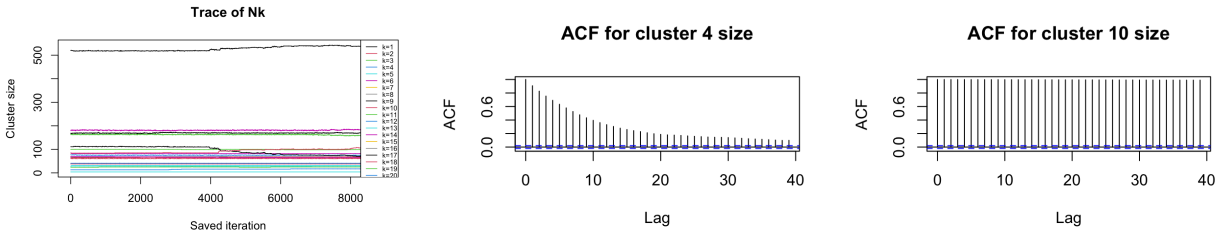
Table 1: Top 10 Brands for Clusters 1, 9, and 19 Identified by the Dirichlet–Multinomial Mixture Model

Cluster 1	**Cluster 9**	**Cluster 19**
fashionnova	anastasiabeverlyhills	liketoknow.it
prettylittlething	morphebrushes	vicidolls
rebelliousfashion	nyxcosmetics	nordstrom
boohoo	benefitcosmetics	target
wilhelminamodels	maccosmetics	guruchoudhary
missyempire	tartecosmetics	liketoknow.it.family
isawitfirst	maybelline	chicwish
missguided	urbandecaycosmetics	sheinofficial
publicdesire	colourpopcosmetics	liketoknow.it.home
misspap	toofaced	danielwellington

- Cluster1 is dominated by fast-fashion accounts such as @fashionnova and *@rebelliousfashion*.
- Cluster 9 is centered on beauty and cosmetic brands such as *@nyxcosmetics* and *@maccosmetics*.
- Cluster 18 is characterized by lifestyle and commerce-related brands such as *@liketoknow.it*, *@nordstrom*, and *@target*.

However, there are also some overlaps between clusters, for example, I found cluster 3 is also dominated by cosmetics account, which has a high similarity as cluster 9.

Diagnostics



(a) Trace of N_k (b) ACF for cluster 4 (c) ACF for cluster 10

Figure 9: Autocorrelation Check

The computed Effective Sample Size (ESS) for 20 clusters are:

Table 2: Effective Sample Sizes (ESS) for Cluster Sizes in the Gibbs Sampler ($K = 20$)

Cluster	1	2	3	4	5	6	7	8	9	10
ESS	0.00	4409.92	2.90	417.74	436.31	436.31	1.14	7.53	15.69	1.13
Cluster	11	12	13	14	15	16	17	18	19	20
ESS	2.47	1.64	68.96	341.39	6.75	9000.00	0.00	59.84	1.71	1.41

Summary statistics:

$$\text{Min} = 0.00, \quad \text{Median} \approx 7.14, \quad \text{Mean} \approx 760.64, \quad \text{Max} = 9000.00.$$

Discussion

Mixture Weights π The clusters with high posterior mixture weight π_k (e.g., 3, 13, 19) showed clear preference patterns, however, we also observe topical overlap among a few active clusters (e.g., two cosmetics components), suggesting possible redundancy. This indicates that only a small number of segments are prevalent in the population.

Cluster Size N_k. The results reveal substantial heterogeneity in the degree of movement across clusters. A number of clusters had little or no change in their assigned sizes; several clusters exhibited moderate fluctuations in their sizes over time, their assignments evolve slowly but still allowing the sampler to explore local modes. This might indicate that only a few clusters are actively used to model meaningful segments.

Next Steps.

- Re-run the sampler with different the number of components K and compare fits (e.g., held-out predictive likelihood).
- Run multiple chains with different initializations; assess convergence using label-invariant summaries (sorted N_k), $\hat{R}$, and ESS.