Chapter 3
Mixture Modeling

Qing Zhou[*,†]

Contents

1. Mixture models

1.1. Definition

Model the distribution of $y = (y_1, y_2, \cdots, y_n)$ as a mixture of K components:

$$\mathbb{P}(y_i|\theta, \lambda) = \sum_{m=1}^{K} \lambda_m f_m(y_i|\theta_m), \tag{1}$$

where λ_m is the proportion of the m^{th} component, $\sum_{m=1}^{K} \lambda_m = 1$, and $f_m(y_i|\theta_m)$ is the distribution of m^{th} component (usually from the same parametric family).

Now let us introduce missing indicator variables $z_i = (z_{i1}, \cdots, z_{iK})$:

$$z_{im} = \begin{cases} 1 & \text{if } y_i \text{ is drawn from the } m^{\text{th}} \text{ mixture component} \\ 0 & \text{otherwise} \end{cases}.$$

Thus, we have the following two-layer model:

$$\begin{aligned} z_i &\sim \mathcal{M}(1, (\lambda_1, \cdots, \lambda_K)), \\ y_i|z_i &\sim f_m(y_i|\theta_m), \text{ if } z_{im} = 1. \end{aligned}$$

[*] UCLA Department of Statistics (email: zhou@stat.ucla.edu).
[†] I thank Elvis Cui for typesetting part of this chapter in LaTex.

It is easy to see that the marginal distribution $[y_i|\theta, \lambda]$ is given by the mixture distribution (1):

$$\mathbb{P}(y_i|\theta, \lambda) = \sum_{z_i} \mathbb{P}(y_i, z_i \mid \lambda, \theta) = \sum_{m=1}^{K} \lambda_m f_m(y_i|\theta_m),$$

by summing over the range of $z_i \in \{e_1, \ldots, e_K\}$, where e_m's are the standard basis vectors in $\mathbb{R}^K$, e.g. $e_1 = (1, 0, \ldots, 0)$.

We may formulate this as a missing data problem:

- $y = (y_1, \cdots, y_n)^\mathsf{T}$: $n \times p$ matrix, observed data (p is the dimension of y_i);
- $z = (z_1, z_2, \cdots, z_n)^\mathsf{T}$: $n \times K$ matrix, missing data.

Write the pdf of $[y_i|z_i]$ as $\prod_{m=1}^{K}(f(y_i|\theta_m))^{z_{im}}$. Therefore, the complete-data likelihood is:

$$\mathbb{P}(y, z|\theta, \lambda) = \prod_{i=1}^{n} \prod_{m=1}^{K} \left(\lambda_m f(y_i|\theta_m)\right)^{z_{im}}.$$

Remark 1. For a distribution $\mathbb{P}_\theta = \mathbb{P}(x|\theta)$, the parameter θ is *identifiable* if the mapping $\theta \mapsto \mathbb{P}_\theta$ is one-to-one. For $\mathbb{P}(y|\theta, \lambda)$ in (1), the parameters (θ, λ) are *not* identifiable due to permutation of the group labels $\{1, \ldots, K\}$. However, the non-identifiability of a mixture model is usually not an issue in practice, since most methods will produce an estimate of the parameters defined by an arbitrary permutation of the group labels.

1.2. MLE by the EM

Log-likelihood of complete data:

$$\log(\mathbb{P}(y, z|\theta, \lambda)) = \sum_{i=1}^{n} \sum_{m=1}^{K} z_{im}[\log \lambda_m + \log f(y_i|\theta_m)].$$

Taking expection w.r.t. $[z|y, \theta^{(t)}, \lambda^{(t)}]$:

$$\begin{aligned} &\mathbb{E}\left[\log(\mathbb{P}(y, z|\theta, \lambda))|y, \theta^{(t)}, \lambda^{(t)}\right] \\ &\quad = \sum_{i=1}^{n} \sum_{m=1}^{K} \mathbb{E}(z_{im}|y_i, \theta^{(t)}, \lambda^{(t)})[\log \lambda_m + \log f(y_i|\theta_m)]. \end{aligned}$$

Calculate the conditional expectation:

$$\begin{aligned} \mathbb{E}(z_{im}|y_i, \theta^{(t)}, \lambda^{(t)}) &= \mathbb{P}(z_{im} = 1|y_i, \theta^{(t)}, \lambda^{(t)}) \\ &= \frac{\mathbb{P}(y_i|z_{im} = 1, \theta_m^{(t)})\mathbb{P}(z_{im=1}|\lambda^{(t)})}{\sum_{j=1}^{K} \mathbb{P}(y_i|z_{ij} = 1, \theta_j^{(t)})\mathbb{P}(z_{ij=1}|\lambda^{(t)})} \\ &= \frac{\lambda_m^{(t)} f(y_i|\theta_m^{(t)})}{\sum_{j=1}^{K} \lambda_j^{(t)} f(y_i|\theta_j^{(t)})} \\ &\triangleq w_{im}^{(t)} : \text{weight of } y_i \text{ from } f(\cdot|\theta_m). \end{aligned}$$

Note that $\sum_m w_{im}^{(t)} = 1$. The $w_{im} = \mathbb{P}(z_{im} = 1 \mid y_i)$ are *posterior* probabilities of $z_{im} = 1$, while $\lambda_m = \mathbb{P}(z_{im} = 1)$ are *prior* probabilities.

Thus, given $(\lambda^{(t)}, \theta^{(t)})$, one iteration if the EM algorithm can be described as follow:

- E-step: Calculate the weights $w_{im}^{(t)}$ for $m = 1, \ldots, K$ and $i = 1, \ldots, n$. Then

$$\begin{aligned} Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) &= \mathbb{E}\left[\log(\mathbb{P}(y, z|\theta, \lambda))|y, \theta^{(t)}, \lambda^{(t)}\right] \\ &= \sum_{m=1}^{K} \left\{ \underbrace{\left(\sum_{i=1}^{n} w_{im}^{(t)}\right)}_{\triangleq w_{\cdot m}^{(t)}} \log \lambda_m + \left(\sum_{i=1}^{n} w_{im}^{(t)} \log f(y_i|\theta_m)\right) \right\} \\ &= \sum_{m=1}^{K} w_{\cdot m}^{(t)} \log \lambda_m + \sum_{m=1}^{K} \left[\sum_{i=1}^{n} w_{im}^{(t)} \log f(y_i|\theta_m)\right]. \end{aligned}$$

- M-step: Let

$$w_{\cdot\cdot}^{(t)} \triangleq \sum_{m=1}^{K} w_{\cdot m}^{(t)} = n,$$

$$Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)}) \triangleq \sum_{i=1}^{n} w_{im}^{(t)} \log f(y_i|\theta_m).$$

Then, for $m = 1, \ldots, K$,

$$\lambda_m^{(t+1)} = \frac{w_{\cdot m}^{(t)}}{w_{\cdot\cdot}^{(t)}} = \frac{w_{\cdot m}^{(t)}}{n}; \tag{2}$$

$$\theta_m^{(t+1)} = \arg\max_{\theta} Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)}). \tag{3}$$

The update of λ by (2) is the same for all models. In the following examples, we show how to update θ_m.

Example 1 (Mixture exponential)**.** Assumptions:

$$y_i|(z_{im} = 1, \theta_m) \sim \mathcal{E}(\theta_m),$$

$$f(y_i|z_{im} = 1, \theta_m) = \frac{1}{\theta_m} \exp\left(-\frac{y_i}{\theta_m}\right).$$

Thus $\mathbb{E}(y_i \mid z_{im} = 1, \theta_m) = \theta_m$.

Calculating Q function:

$$\begin{aligned} Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)}) &= \sum_{i=1}^{n} w_{im}^{(t)} \log\left[\frac{1}{\theta_m} \exp\left(-\frac{y_i}{\theta_m}\right)\right] \\ &= -w_{\cdot m}^{(t)} \log \theta_m - \frac{\sum_{i=1}^{n} w_{im}^{(t)} y_i}{\theta_m}. \end{aligned}$$

Taking derivative and set it to zero:

$$\frac{\partial Q_m}{\partial \theta_m} = 0 \Rightarrow \theta_m^{(t+1)} = \frac{\sum_{i=1}^n w_{im}^{(t)} y_i}{w_{\cdot m}^{(t)}},$$

which is a *weighted average* of y_i.

Example 2 (Exponential family)**.** Suppose

$$f(y_i|\theta_m, z_{im} = 1) = h(y_i)c(\theta_m)\exp[\phi(\theta_m)^\mathsf{T} t(y_i)], \qquad m = 1, \ldots, K.$$

- E-step:

$$\begin{aligned} Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)}) &= \sum_{i=1}^n w_{im}^{(t)} \left[\log h(y_i) + \log c(\theta_m) + \phi(\theta_m)^\mathsf{T} t(y_i)\right] \\ &= w_{\cdot m}^{(t)} \log c(\theta_m) + \phi(\theta_m)^\mathsf{T} \left(\sum_{i=1}^n w_{im}^{(t)} t(y_i)\right) + \text{const.} \end{aligned}$$

- M-step: ${\theta_m}^{(t+1)}$ is the solution (for θ) to

$$\sum_{i=1}^n w_{im}^{(t)} t(y_i) = \mathbb{E}_{\theta_m}\left[\sum_{i=1}^n w_{im}^{(t)} t(y_i)\right] = w_{\cdot m}^{(t)} \mathbb{E}_{\theta_m}[t(y_1)].$$

  Remark: Compare to complete data, where $\widehat{\theta}_{\text{MLE}}$ satisfies

$$\sum_{i=1}^n t(y_i) = n\mathbb{E}_\theta[t(y_1)].$$

2. Model-based clustering

Clustering problem: Suppose we observe $y_1, \cdots, y_n$ ($y_i \in \mathbb{R}^p$) from K groups. Now we want to group them into K clusters. This problem can be illustrated by Figure 1.

Assumptions: Denote by z_i as the cluster label of y_i, which is hidden (or latent variable).

$$z_i \sim \mathcal{M}(1, \lambda), \quad \lambda = (\lambda_1, \cdots, \lambda_K)$$
$$y_i|z_{im} = 1 \sim \mathcal{N}_p(\mu_m, \Sigma_m).$$

Estimation: We want to find MLE of parameters $\theta = (\lambda, \mu_m, \Sigma_m, m = 1, \ldots, K)$. Then predict cluster label according to $\mathbb{P}(z_{im} = 1|y_i, \widehat{\theta})$.

- E-step: For $i = 1, \ldots, n$ and $m = 1, \ldots, K$, calculate

$$w_{im}^{(t)} = \frac{\lambda_m^{(t)} \phi_p(y_i; \mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{j=1}^K \lambda_j^{(t)} \phi_p(y_i, \mu_j^{(t)}, \Sigma_j^{(t)})}.$$

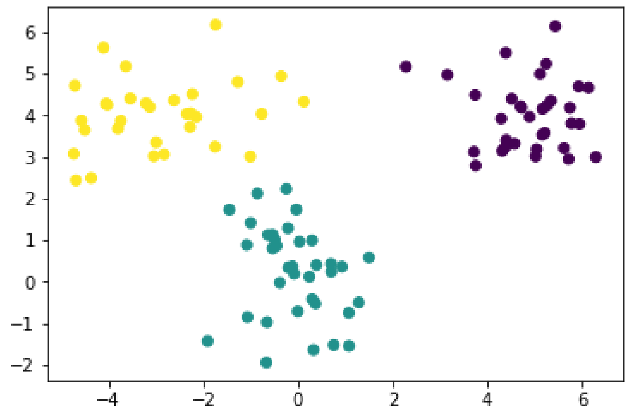

Fig 1: Scatter plot of three clusters of data points.

- M-step: Update $\lambda^{(t+1)}$ by (2). For $m = 1, \cdots, K$, solve

$$\sum_i w_{im}^{(t)} y_i = w_{\cdot m}^{(t)} \mu_m$$
$$\sum_i w_{im}^{(t)} y_i y_i^\mathsf{T} = w_{\cdot m}^{(t)} \left(\Sigma_m + \mu_m \mu_m^\mathsf{T}\right)$$

for μ_m and Σ_m to update

$$\mu_m^{(t+1)} = \frac{\sum_i w_{im}^{(t)} y_i}{w_{\cdot m}^{(t)}},$$
$$\Sigma_m^{(t+1)} = \frac{\sum_i w_{im}^{(t)} y_i y_i^\mathsf{T}}{w_{\cdot m}^{(t)}} - \mu_m^{(t+1)} (\mu_m^{(t+1)})^\mathsf{T}.$$

Prediction: After EM converges, the predicted cluster label

$$\widehat{z}_i = \operatorname*{argmax}_{1 \le m \le K} \mathbb{P}(z_{im} = 1 | y_i, \widehat{\theta}) = \operatorname*{argmax}_{1 \le m \le K} w_{im}^{(T)},$$

where T indexes the last iteration and $\widehat{\theta} = \theta^{(T)}$.

Simplification: When p is big, $\Sigma_m(p \times p)$ has too many parameters, and we may simplify the model by assuming $\Sigma_m = \sigma_m^2 I_p$. This links us to K-means clustering.

Theorem 1. *Assume $\Sigma_1 = \cdots = \Sigma_K = \sigma^2 I_p$, and σ^2 is known. If $\sigma^2 \to 0^+$, then the above EM algorithm is equivalent to K-means clustering.*

Proof. If $\Sigma_m = \sigma^2 I_p$, the E-step simplifies to

$$w_{im}^{(t)} = \frac{\lambda_m^{(t)} \exp\left(-\frac{\|y_i - \mu_m^{(t)}\|_2^2}{2\sigma^2}\right)}{\sum_{j=1}^K \lambda_j^{(t)} \exp\left(-\frac{\|y_i - \mu_j^{(t)}\|_2^2}{2\sigma^2}\right)}.$$

As $\sigma^2 \to 0^+$,

$$w_{im}^{(t)} = \begin{cases} 1 & \text{if } m = \text{argmin}_j \|y_i - \mu_j^{(t)}\|_2^2 \\ 0 & \text{otherwise} \end{cases},$$

i.e., assigning y_i to the closest center. Let $\mathcal{C}_m^{(t)} = \{i : w_{im}^{(t)} = 1\}$ be the m^{th} cluster, and $|\mathcal{C}_m^{(t)}|$ its size, in the current iteration. Then, the updated parameter in the M-step becomes

$$\mu_m^{(t+1)} = \frac{\sum_{i \in \mathcal{C}_m^{(t)}} y_i}{|\mathcal{C}_m^{(t)}|},$$

i.e., update μ_m by the sample mean of $\mathcal{C}_m^{(t)}$. □

3. Motif discovery

3.1. Problem formulation

In genomics and molecular biology, a sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance. Figure 2 illustrates a DNA sequence motif that is recognized by a transcription factor (TF). After the TF binds to the DNA sequence, the downstream gene can be activated or suppressed. Review of sequence motifs and motif finding methods can be found in Jensen et al. (2004).

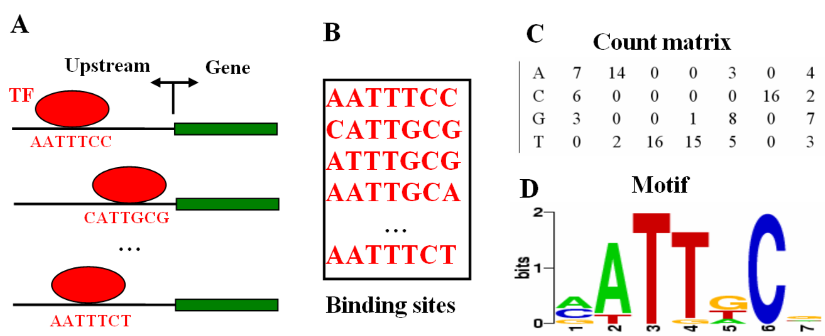

Fig 2: Sequence motif. (A) Upstream sequences of genes that share a common motif recognized by a TF. (B) Examples of the TF binding sites (motif sequences). (C) Count matrix from the motif sequences. (D) Logo plot for the motif.

Given a set of sequences, we want to identify the motif sites in these sequences. This is the motif finding problem (Figure 3), which can be formulated as a mixture model with two component distributions:

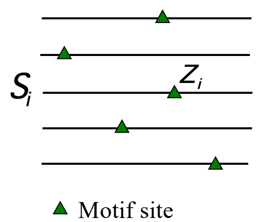

Fig 3: Motif finding problem, one motif site (at Z_i) in each sequence S_i.

Observed Data	Missing Data	Parameters
$S = (S_1, \cdots, S_n)$	$Z = (Z_1, \cdots, Z_n)$	Θ: motif pattern, θ_0: background

- $S = (S_1, S_2, \ldots, S_n)$: sequences on alphabet $\{A, C, G, T\}$ (observed data).
- $Z = (Z_1, Z_2, \cdots, Z_n)$: motif site locations, that is, Z_i is the beginning location of the motif site in S_i. Z is unobserved (missing data).
- **Motif model**: $X = (x_1, \cdots, x_w)$, motif of length w, $x_i \in \{A, C, G, T\}$ and $x_i \perp x_j$. Each component x_i of X follows a multinomial distribution with unknown parameter $\theta_i = (\theta_{iA}, \theta_{iC}, \theta_{iG}, \theta_{iT})$. Thus, X can be viewed as a $4 \times w$ counting (indicator) matrix.
  Put $\Theta = (\theta_1, \cdots, \theta_w)$: unknown parameters. The distribution of X is a *product multinomial* distribution with parameter Θ.
  Example: $\mathbb{P}\{X = (AATGC)|\Theta\} = \theta_{1A}\theta_{2A}\theta_{3T}\theta_{4G}\theta_{5C}$.
- **Background model**: $\widetilde{x} \sim_{iid} \mathcal{M}(\theta_0)$, $\theta_0 = (\theta_{0A}, \theta_{0C}, \theta_{0G}, \theta_{0T})$. That is, $\mathbb{P}(\widetilde{x} = j|\theta_0) = \theta_{0j},\ j \in \{A, C, G, T\}$. Assume that θ_0 is known.

3.2. Maximum likelihood via EM

Define

$$S_i(j, w) := \text{the segment of } S_i \text{ starting at } j^{\text{th}} \text{ position with length } w.$$

Note that j ranges from 1 to $\ell_i := L_i - w + 1$, where $L_i = |S_i|$ (total length of i^{th} sequene). The MLE for Θ is given by

$$\begin{aligned}\widehat{\Theta} &= \underset{\Theta}{\operatorname{argmax}} \underbrace{\mathbb{P}(S|\Theta)}_{\text{obs-data lik}} \\ &= \underset{\Theta}{\operatorname{argmax}} \sum_Z \underbrace{\mathbb{P}(S, Z|\Theta)}_{\text{comp-data lik}} .\end{aligned}$$

Now look at one sequence (recall that background model θ_0 is known). Assume Z_i is uniform *in priori*:

$$\begin{aligned}
\mathbb{P}(S_i, Z_i = j|\Theta) &= \mathbb{P}(Z_i = j)\mathbb{P}(S_i|Z_i = j, \Theta) \\
&= \frac{1}{\ell_i}\mathbb{P}(S_i(j,w)|\Theta)\mathbb{P}(S_i \setminus S_i(j,w)|\theta_0) \\
&= \frac{1}{\ell_i}\frac{\mathbb{P}(S_i(j,w)|\Theta)}{\mathbb{P}(S_i(j,w)|\theta_0)}\mathbb{P}(S_i|\theta_0) \\
&\propto \frac{\mathbb{P}(S_i(j,w)|\Theta)}{\mathbb{P}(S_i(j,w)|\theta_0)} \equiv r_{ij}(\Theta) \quad \text{(likelihood ratio)}. \qquad (4)
\end{aligned}$$

Therefore, the posterior probability of $[Z_i = j|S_i]$ is

$$w_{ij}(\Theta) := \mathbb{P}(Z_i = j|S_i, \Theta) = \frac{\mathbb{P}(S_i, Z_i = j|\Theta)}{\sum_{k=1}^{\ell_i}\mathbb{P}(S_i, Z_i = k|\Theta)} = \frac{r_{ij}(\Theta)}{\sum_{k=1}^{\ell_i} r_{ik}(\Theta)}.$$

Since $[S_i|Z_i]$ is an exponential family (product multinomial), as long as we derive the sufficient statistic for Θ, it can be used to compute MLE for Θ by the EM algorithm. (Recall the EM for exponential families).

The count matrix is a sufficient statistic for Θ. For example,

$$S_i(j,w) = A \;\; C \;\; C \;\; T \;\; G$$

$$C(S_i(j,w)) := \begin{array}{c} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{array}{l} A \\ C \\ G \\ T \end{array}.$$

If $X_1, \cdots, X_n$ are the count matrices of the n motif sequences (Z known), then the MLE of Θ is

$$\widehat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} X_i := \frac{1}{n}X. \qquad (5)$$

For example, the (total) count matrix of $n = 15$ motif sequences

$$X := \sum_{i=1}^{n} X_i = \begin{bmatrix} 10 & 1 & \cdots & 1 \\ 1 & 10 & \cdots & 2 \\ 1 & 3 & \cdots & 7 \\ 3 & 1 & \cdots & 5 \end{bmatrix}_{4\times w}.$$

Thus, the EM algorithm can be done by iterating between:

- (E-step) Given $\Theta^{(t)}$, find $\mathbb{E}(X|S,\Theta^{(t)})$:

$$\begin{aligned}
\mathbb{E}(X_i|S_i,\Theta^{(t)}) &= \sum_{j=1}^{\ell_i} C[S_i(j,w)]\mathbb{P}(Z_i=j|S_i,\Theta^{(t)}) \\
&= \sum_{j=1}^{\ell_i} w_{ij}(\Theta^{(t)})C[S_i(j,w)], \\
\Longrightarrow X^{(t)} &:= \mathbb{E}(X|S,\Theta^{(t)}) \\
&= \sum_{i=1}^{n} \mathbb{E}(X_i|S_i,\Theta^{(t)}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{\ell_i} w_{ij}(\Theta^{(t)})C[S_i(j,w)].
\end{aligned}$$

- (M-step) Regarding $X^{(t)}$ as the sufficient statistic, find MLE as in (5):

$$\Theta^{(t+1)} = \frac{X^{(t)}}{n}.$$

3.3. Bayesian inference via Gibbs sampler

Assume a conjugate prior $\theta_j \sim \mathrm{Dir}(\alpha,\dots,\alpha)$ independently for $j=1,\dots,w$. In short, we say the prior of Θ is product-Dirichlet,

$$\Theta \sim \text{Prod-Dir}(\alpha). \tag{6}$$

Let $X_{\bullet j}$ be the j^{th} column of the count matrix X: $X_{\bullet j} \mid \theta_j \sim M(n,\theta_j)$. Then the posterior distribution

$$\theta_j \mid X_{\bullet j} \sim \mathrm{Dir}(X_{\bullet j}+\alpha),\ j=1,\dots,w \iff \Theta \mid X \sim \text{Prod-Dir}(X+\alpha). \tag{7}$$

The posterior mean

$$\mathbb{E}(\Theta \mid X) = \frac{X+\alpha}{n+4\alpha}.$$

Under this prior, we develop a Gibbs sampler to draw $[Z_1,\dots,Z_n \mid S]$ to predict the motif locations. The Gibbs sampler cycles through conditional distributions $[Z_i \mid Z_{-i}, S]$ for $i=1,\dots,n$ in each iteration. The key is to calculate

$$\begin{aligned}
\mathbb{P}(Z_i=j \mid Z_{-i},S) &\propto \mathbb{P}(S_i, Z_i=j \mid Z_{-i}, S_{-i}) \\
&= \int_\Theta \mathbb{P}(S_i, Z_i=j \mid \Theta)\, p(\Theta \mid X_{-i}) d\Theta,
\end{aligned}$$

where the count matrix X_{-i} is computed from (S_{-i}, Z_{-i}) and the posterior distribution $\Theta \mid X_{-i} \sim \text{Prod-Dir}(X_{-i} + \alpha)$ as in (7). Plugging (4),

$$\mathbb{P}(Z_i = j \mid Z_{-i}, S) \propto \int_\Theta \frac{\mathbb{P}(S_i(j,w)|\Theta)}{\mathbb{P}(S_i(j,w)|\theta_0)} p(\Theta \mid X_{-i}) d\Theta = r_{ij}(\widehat{\Theta}_{-i}),$$

where $\widehat{\Theta}_{-i}$ is the posterior mean

$$\widehat{\Theta}_{-i} = \mathbb{E}(\Theta \mid X_{-i}) = \frac{X_{-i} + \alpha}{n - 1 + 4\alpha}. \tag{8}$$

After normalization,

$$\mathbb{P}(Z_i = j \mid Z_{-i}, S) = \frac{r_{ij}(\widehat{\Theta}_{-i})}{\sum_{k=1}^{\ell_i} r_{ik}(\widehat{\Theta}_{-i})} = w_{ij}(\widehat{\Theta}_{-i}), \qquad j = 1, \ldots, \ell_i. \tag{9}$$

In summary, each iteration of this Gibbs sampler consists of the following loop.
For $i = 1, \ldots, n$:

1. Compute the posterior mean $\widehat{\Theta}_{-i} = \mathbb{E}(\Theta \mid X_{-i})$ by (8).
2. Draw $[Z_i \mid Z_{-i}, S]$ according to (9).

4. Problem set

Datasets can be downloaded from the course site.

1. Suppose that X follows a two-component mixture distribution with mixture proportions λ_1 and λ_2 ($\lambda_1 + \lambda_2 = 1$). The mean and the variance of the m^{th} component distribution are μ_m and σ_m^2, respectively, for $m = 1, 2$. Find $\mathbb{E}(X)$ and $\text{Var}(X)$.
2. Dataset 1 consists of data points from three clusters. Suppose the data points in the mth ($m = 1, 2, 3$) cluster are iid from $\mathcal{N}_p(\mu_m, \sigma_m^2 I_p)$, where $p = 2$ is the dimension of the data.
   (a) Derive an EM algorithm to find the MLE of the unknown parameters.
   (b) Implement the EM algorithm to cluster these data points into three groups. Report the estimated parameters and make a scatterplot of the data points with your predicted cluster labels.
3. We have observed $n = 10$ sites of a motif, summarized into a count matrix X_{obs} shown in Table 1. A position-specific weight matrix Θ is used as the model for the motif sites. Assume an known iid background model with $\theta_0 = (0.24, 0.26, 0.26, 0.24)$ for $\{A, C, G, T\}$. In addition to X_{obs}, we know that the sequence

   $$S = \text{ACCATTATCCCTGT}$$

   contains another site of this motif and let $Z \in \{1, \ldots, 10\}$ be its start position. Assume that the marginal probability $\mathbb{P}(Z = i)$ is identical for all possible i.

TABLE 1
Observed count matrix X_{obs}

Position	1	2	3	4	5
A	1	9	0	0	8
C	3	0	0	0	0
G	6	1	0	0	1
T	0	0	10	10	1

(a) Let $\hat{\Theta}_{\text{obs}} = \frac{1}{n+4\alpha}(X_{\text{obs}} + \alpha)$, where $\alpha = 1$ is a pseudo count. Find the most likely start position of the site in S by

$$\max_{1 \leq i \leq 10} \mathbb{P}(Z = i \mid S, \hat{\Theta}_{\text{obs}}).$$

(b) Regarding both X_{obs} and S as our data, develop a method to find the MLE of Θ, i.e.,

$$\hat{\Theta}_{\text{MLE}} = \operatorname*{argmax}_{\Theta} \mathbb{P}(S, X_{\text{obs}} \mid \Theta).$$

Implementation is not required. Just write down the main steps.

(c) Hereafter, we consider this problem in a Bayesian way, assuming a Product-Dirichlet prior for Θ as in (6) with $\alpha = 1$. Find the posterior distribution of Θ given the observed count matrix, $[\Theta \mid X_{\text{obs}}]$.

(d) Implement a Monte Carlo method to draw 2000 samples of Θ from the posterior distribution $[\Theta \mid X_{\text{obs}}, S]$. Use the samples to approximate the posterior mean $\widehat{\Theta}_B = \mathbb{E}[\Theta \mid X_{\text{obs}}, S]$ and the posterior probabilities $\mathbb{P}(\theta_{jk} > 0.5 \mid X_{\text{obs}}, S)$ for $j = 1, \ldots, 5$ and $k \in \{A, C, G, T\}$.
Hint:

$$p(\Theta \mid X_{\text{obs}}, S) = \sum_i p(\Theta \mid X_{\text{obs}}, S, Z = i)\mathbb{P}(Z = i \mid X_{\text{obs}}, S).$$

References

JENSEN, S. T., LIU, X. S., ZHOU, Q. and LIU, J. S. (2004). Computational discovery of gene regulatory binding motifs: A Bayesian perspective. *Statistical Science* **19** 188-204.