Chapter 2
Mixture Modeling

Qing Zhou∗†

Contents

1 Mixture models .................................................. 1
  1.1 Definition .................................................. 1
  1.2 MLE by the EM ............................................. 2
2 Model-based clustering ........................................... 4
3 Motif discovery .................................................. 6
  3.1 Problem formulation ....................................... 6
  3.2 Maximum likelihood via EM ................................ 7
4 Problem set .................................................... 9
References ..................................................... 10

1. Mixture models

1.1. Definition

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

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

where \( \lambda_m \) 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 \( Z_i = (z_{i1}, \cdots, z_{iK}) \):

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

Thus we have the following two-layer model:

\[
Z_i \sim M(1, (\lambda_1, \cdots, \lambda_K)),
\]

\[
Y_i|Z_i \sim f(y_i|\theta_m), \text{if } z_{im} = 1.
\]

\( y = (y_1, \cdots, y_n)^T \): \( n \times p \) matrix, observed data (\( p \) is the dimensional of \( y_i \)).

∗UCLA Department of Statistics (email: zhou@stat.ucla.edu).
†I thank Elvis Cui for typesetting part of this chapter in Latex.
\( z = (z_1, z_2, \cdots, z_n)^T \): \( n \times K \) matrix, missing data.

Also we can 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:

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

Note that the marginal distribution of \( y_i \) is identical to (1):

\[
P(y_i | \theta, \lambda) = \sum_{z_i} \prod_{m=1}^K (\lambda_m f(y_i | \theta_m))^{z_{im}} = \sum_{m=1}^K \lambda_m f_m(y_i | \theta_m),
\]

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. \( e_1 = (1, 0, \ldots, 0) \).

**Remark 1.** Parameters are **NOT** identifiable if the same likelihood function is obtained for more than one choice of the model parameters. In \( P(Y | \theta, \lambda) \), the parameters \((\theta, \lambda)\) are **not identifiable** due to permutation of the group labels. However, the non-identifiability issue can be avoided by restricting to a specific ordering among the groups.

### 1.2. MLE by the EM

Log-likelihood of complete data:

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

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

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

Calculate the conditional expectation:

\[
E(z_{im} | y, \theta^{(t)}, \lambda^{(t)}) = P(z_{im} = 1 | y, \theta^{(t)}, \lambda^{(t)})
\]

\[
= \frac{P(y_i | z_{im} = 1, \theta_m^{(t)}) P(z_{im} = 1 | \lambda^{(t)})}{\sum_{j=1}^K P(y_i | z_{ij} = 1, \theta_j^{(t)}) P(z_{ij} = 1 | \lambda^{(t)})}
\]

\[
= \frac{\lambda_m^{(t)} f(y_i | \theta_m^{(t)})}{\sum_{j=1}^K \lambda_j^{(t)} f(y_i | \theta_j^{(t)})}
\]

\( \triangleq w_{im}^{(t)} \): weight of \( y_i \) from \( f(\cdot | \theta_m) \).
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

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

- **M-step**: Let

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

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

  Then

  $\lambda_m^{(t+1)} = \frac{w_{m}^{(t)}}{w^{(t)}} = \frac{w_{m}^{(t)}}{n},$

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

**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:

$Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)}) = \sum_{i=1}^{n} w_{im}^{(t)} \log \left[ \frac{1}{\theta_m} \exp \left( -\frac{y_i}{\theta_m} \right) \right]
= -w_{m}^{(t)} \log \theta_m - \frac{\sum_{i=1}^{n} w_{im}^{(t)} y_i}{\theta_m}.$
Taking derivative and set it to zero:

\[
\frac{\partial Q_m}{\partial \theta_m} = 0 \Rightarrow \theta_m^{(t+1)} = \frac{\sum_{i=1}^n w_{im}^{(t)} y_i}{w_m^{(t)}},
\]

which is a weighted average of \( y_i \).

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

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

- **E-step:**

\[
Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)}) = \sum_{i=1}^n w_{im}^{(t)} \left[ \log h(y_i) + \log c(\theta_m) + \phi(\theta_m)^{T}t(y_i) \right]
\]

\[
= w_m^{(t)} \log c(\theta_m) + \phi(\theta_m)^{T} \left( \sum_{i=1}^n w_{im}^{(t)}t(y_i) \right) + \text{const}.
\]

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

\[
\sum_{i=1}^n w_{im}^{(t)}t(y_i) = E_{\theta}\left[ \sum_{i=1}^n w_{im}^{(t)}t(y_i) \right] = w_m^{(t)}E_{\theta}[t(y_1)].
\]

Remark: Compare to complete data, where \( \hat{\theta}_{\text{MLE}} \) satisfies

\[
\sum_{i=1}^n t(y_i) = nE_{\theta}[t(y_1)].
\]

2. **Model-based clustering**

*Clustering problem:* Suppose we observe \( y_1, \ldots, y_n \ (y_i \in \mathbb{R}^p) \) 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).

\[
z_i \sim \mathcal{M}(1, \lambda), \quad \lambda = (\lambda_1, \ldots, \lambda_K)
\]

\[
y_i|z_{im} = 1 \sim \mathcal{N}(\mu_m, \Sigma_m).
\]

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

- **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:** For $m = 1, \cdots, K$, solve
  
  \[
  \sum_i w_{im}^{(t)} y_i = w_m^{(t)} \mu_m \\
  \sum_i w_{im}^{(t)} y_i y_i^T = w_m^{(t)} (\Sigma_m + \mu_m \mu_m^T)
  \]
  
  for $\mu_m$ and $\Sigma_m$ to update
  
  \[
  \mu_m^{(t+1)} = \frac{\sum_i w_{im}^{(t)} y_i}{w_m^{(t)}} \\
  \Sigma_m^{(t+1)} = \frac{\sum_i w_{im}^{(t)} y_i y_i^T}{w_m^{(t)}} - \mu_m^{(t+1)} (\mu_m^{(t+1)})^T.
  \]

  **Prediction:** After EM converges, the predicted cluster label
  
  \[
  \hat{z}_i = \arg\max_{1 \leq m \leq K} P(z_{im} = 1 | y_i, \hat{\theta}) = \arg\max_{1 \leq m \leq K} w_{im}^{(T)}
  \]
  
  where $T$ indexes the last iteration and $\hat{\theta} = \theta^{(T)}$.

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

**Theorem 1.** Assume $\Sigma_1 = \cdots = \Sigma_K = \sigma^2 I_p$, and $\sigma^2$ 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 \( \sigma^2 \to 0^+ \),
\[
w_{im}^{(t)} = \begin{cases} 
1 & \text{if } m = \arg\min \| y_i - \mu_j^{(t)} \|_2 \\
0 & \text{otherwise}
\end{cases},
\]
i.e., assigning \( y_i \) to the closest center. Let \( c_m^{(t)} = \{ i : w_{im}^{(t)} = 1 \} \) be the \( m \)th cluster, and \( |c_m^{(t)}| \) its size, in the current iteration. Then, the updated parameter in the M-step becomes
\[
\mu_m^{(t+1)} = \frac{\sum_{i \in c_m^{(t)}} y_i}{|c_m^{(t)}|},
\]
i.e., update \( \mu_m \) by the sample mean of \( c_m^{(t)} \).

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 $S_i$.

<table>
<thead>
<tr>
<th>Observed Data</th>
<th>Missing Data</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S = (S_1, \ldots, S_n)$</td>
<td>$Z = (Z_1, \ldots, Z_n)$</td>
<td>$\Theta$: motif pattern, $\theta_0$: background</td>
</tr>
</tbody>
</table>

- $S = (S_1, S_2, \ldots, S_n)$: sequences on alphabet $\{A, C, G, T\}$ (observed data).
- $Z = (Z_1, Z_2, \ldots, Z_n)$: 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:** $X = (x_1, \ldots, x_w)$, 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 $\Theta = (\theta_1, \ldots, \theta_w)$: unknown parameters. The distribution of $X$ is a **product multinomial** distribution with parameter $\Theta$.

Example: $P\{X = (AATGC)|\Theta\} = \theta_{1A}\theta_{2A}\theta_{3T}\theta_{4G}\theta_{5C}$.

- **Background model:** $\tilde{x} \sim_{iid} M(\theta_0)$, $\theta_0 = (\theta_{0A}, \theta_{0C}, \theta_{0G}, \theta_{0T})$. That is, $P(\tilde{x} = j|\theta_0) = \theta_{0j}$, $j \in \{A, C, G, T\}$. Assume that $\theta_0$ is known.

### 3.2. Maximum likelihood via EM

Define

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

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

$$\hat{\Theta} = \arg\max_{\Theta} \prod_{i} P(S_i|\Theta) = \arg\max_{\Theta} \sum_{Z} P(S, Z|\Theta).$$

Now look at one sequence (recall that background model $\theta_0$ is known). As-
sume $Z_i$ is uniform in priori:

$$
P(S_i, Z_i = j | \Theta) = \frac{1}{l_i} P(S_i | \Theta) P(S_i \setminus S_i(j,w) | \theta_0) = \frac{1}{l_i} \frac{P(S_i(j,w) | \Theta)}{P(S_i | \theta_0)} P(S_i | \theta_0)
$$

\[ \propto \frac{P(S_i(j,w) | \Theta)}{P(S_i(j,w) | \theta_0)} \equiv r_{ij}(\Theta) \text{ (likelihood ratio)}. \]

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

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

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

The count matrix is a sufficient statistic for $\Theta$. For example,

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

$$C(S_i(j,w)) := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} A \\ C \\ G' \\ T \end{bmatrix}$$

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

$$\hat{\Theta} = \frac{1}{n} \sum_{i=1}^{n} X_i := \frac{1}{n} X. \quad (2)$$

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

\[
\mathbb{E}(X_i|S_i, \Theta^{(t)}) = \sum_{j=1}^{l_i} C[S_i(j, w)]P(Z_i = j|S_i, \Theta^{(t)}) \\
= \sum_{j=1}^{l_i} w_{ij}(\Theta^{(t)})C[S_i(j, w)],
\]

$$\Rightarrow X^{(t)} := \mathbb{E}(X|S, \Theta^{(t)}) \quad \text{with} \quad X^{(t)} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i|S_i, \Theta^{(t)}) \quad \text{and} \quad \sum_{j=1}^{n} \sum_{i=1}^{l_i} w_{ij}(\Theta^{(t)})C[S_i(j, w)].$$

• (M-step) Regarding $X^{(t)}$ as the sufficient statistic, find MLE as in (2):

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

4. Problem set

Datasets can be downloaded from the CCLE course site.

1. Suppose that $X$ follows a two-component mixture distribution with mixture proportions $\lambda_1$ and $\lambda_2$ ($\lambda_1 + \lambda_2 = 1$). The mean and the variance of the $m^{th}$ component distribution are $\mu_m$ and $\sigma^2_m$, 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 $m^{th}$ ($m = 1, 2, 3$) cluster are iid from $\mathcal{N}_p(\mu_m, \sigma^2_m I_p)$, 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 $X_{\text{obs}}$ shown in Table 1. A position-specific weight matrix $\Theta$ is used as the

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>
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 \( X_{\text{obs}} \), 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 \( P(Z = i) \) is identical for all possible \( i \).

(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

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

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

\[
\hat{\Theta}_{\text{MLE}} = \arg \max_{\Theta} P(S, X_{\text{obs}} \mid \Theta).
\]

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

References