EM Clustering in Exponential Families

Elvis Cui

MS in Biostatistics, UCLA

October 28, 2019

Overview

Book chapters, papers and monograph

- Maximum Likelihood from Incomplete Data via the EM Algorithm, Dempster et al., JRSSB, 1977. [paper]
- The EM Algorithm and Extensions, G. McLachlan, Wiley, 2008. [monograph] [I didn't read it]
- Pattern recognition and machine learning, C. Bishop, Springer, 2006. [Book chapter 9]
- Elements of statistical learning, R. Tibshirani et al, Springer, 2009. [Book chapter 8]
- Modern multivariate statistical techniques, A. Izeman, Springer, 2013. [Book chapter 12]
- Mathematical statistics: basic ideas and selected topics(Volume I). Bickel and Doskum, Chapman, 2001. [Book chapter 2]
- Theory of point estimation, E. Lehmann, 1998. [Book section 6.4]

Other clustering methods

- Hierarchical clustering (a class of methods)
- Partitioning methods
- Self-Organizing Maps (a.k.a. SOMs)
- Block Clustering
- Mixture Models

We will focus on mixture models and EM clustering. Note that $K-$means can be derived as a special case of Gaussian Mixture Model.

Expectation-Maximization (EM): Structure of Dataset

- **Complete data set**: (Y_{obs}, Y_{miss}) (missing data problem in biostatistics) or (Y, Z) (clustering model, hidden Markov model, etc.)
- **Observed data likelihood**: $\mathbb{P}(Y_{obs}|\theta) = \int \mathbb{P}(Y_{obs}, Y_{miss|\theta})dY_{miss}$
- **Maximum likelihood estimation**: $\hat{\theta}_{MLE} := \arg\max_{\theta\in\Theta} \mathbb{P}(Y_{obs}|\theta)$

1. **Problem of deriving MLE directly**: $\hat{\theta}_{MLE}$ is in general difficult to solve (optimize w.r.t. a non-convex function or computationally expensive).
2. **Solution I**: Gradient descent (works well, but we will not dicuss it) (Bishop, 2006).
3. **Solution II**: EM algorithm.

EM Algorithm for parameter estimation

Definition 1 (EM Algorithm)

Suppose we are interested in estimating θ. First, we start with randomly initiated $\theta^{(0)}$. For the $(t+1)^{th}$ iteration, do:

- E-step: calculate conditional expectation:

$$Q(\theta|\theta^t) := \mathbb{E}\left[\log \mathbb{P}(Y_{obs}, Y_{miss}|\theta) | Y_{obs}, \theta^{(t)}\right]$$

- M-step: update parameter:

$$\theta^{(t+1)} := \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(t)})$$

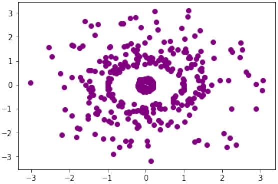
- Iterate the above 2 steps until further notice.

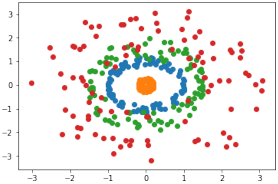
Sidenotes for EM

- If either E or M step is difficult, then EM algorithm is not suitable (Bickel and Doskum, 2001).
- $\log \mathbb{P}(Y_{obs}|\theta)$ will increase after each iteration (for a proof, see, e.g. Lehmann, 1998 or the original paper section 2). Thus, we are actually maximizing likelihood function at each iteration.
- A MCMC version of EM algorithm can be modified (R. Tibshirani et al, 2009).
- EM algorithm can be considered as a special case of MM algorithm (Maximization-minimization).
- The convergence of the EM algorithm is proved in 1983 by a student of Peter Bickel (Jeff Wu).

Finite Mixture Model Set-up

Given the picture in the left, how many clusters you will predict and how you will assign each data point to these cluters ? (The right hand side is the ground truth).





Model Set-up: Marginal distribution

In the original paper (Dempster et al), they didn't give a clear formulation of finite mixture models although it is mentioned in section 3.4.

- **Marginal distribution:** We regard our data set $y = (y_1, \cdots, y_n)$ as a mixture of K components. i.e.:

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

1. y_i: a random variable or vector.
2. λ_m: the proportion of the population from the m^{th} component, $\sum_{m=1}^{K} \lambda_m = 1$.
3. $f_m(y_i|\theta_m)$: pdf or pmf of m^{th} component (can be in the same parametric family).

Model set-up: missing variables or indicators

- **Missing indicator (unobserved)**: $Z_i = (z_{i1}, \cdots, z_{ik})$:

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

  Note that $z_{im} = 1$ with probability λ_m (Why?).

- According to Izenmann, such indicator "was a key innovation of Dempster, Laird and Rubin, 1977".
- **Conditional distribution**:

$$Z_i | \lambda \sim \mathcal{M}(1, (\lambda_1, \cdots, \lambda_K))$$

$$Y_i | Z_i \sim f_m(y_i | \theta_m), \text{if } z_{im} = 1$$

  Where $\mathcal{M}$ denotes multinomial distribution.

Joint likelihood function

Note: a.k.a. complete-data likelihood.

- **Conditional distribution**: $\mathbb{P}(Y_i|Z_i, \theta) = \prod_{m=1}^{K} (f_m(y_i|\theta_m)^{z_{im}})$
- **Complete-data likelihood** (n data points in total):

$$\mathbb{P}(Y, Z|\theta, \lambda) = \prod_{i=1}^{n} \prod_{m=1}^{K} (\lambda_m f(y_i|\theta_m))^{z_{im}}$$

- EM algorithm requires us to compute the expectation of log-joint likelihood function given Y, θ and λ.
- Parameters to be estimated: $\theta = (\theta_1, \cdots, \theta_K), \lambda = (\lambda_1, \cdots, \lambda_K)$
- Observed data: $Y \in \mathbb{R}^{p \times n}$
- Missing data (to be predicted): $Z \in \mathbb{R}^{K \times n}$

MLE via EM algorithm

- Step 1: Derive log-likelihood:

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

- Step 2: Take expectation w.r.t. $Z|y, \theta^{(t)}, \lambda^{(t)}$ (since $\mathbb{E}(y_i|y) = y_i$):

$$\mathbb{E}\left[\log(\mathbb{P}(Y, Z|\theta, \lambda))|y, \theta^{(t)}, \lambda^{(T)}\right] =$$

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

- How to compute $\mathbb{E}(z_{im}|y, \theta^{(t)}, \lambda^{(t)})$?

MLE via EM algorithm

- Recall Bayes' theorem, we have:

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

- Step 3: Maximize $Q(\theta|\theta^{(t)})$ (conditional expectation) to get $\hat{\theta}$ and $\hat{\lambda}$.

EM algorithm for finite mixture model

Definition 2

- E-step

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

EM algorithm for finite mixture model

Definition 3 (continued)

- M-step: First define:

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

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

Then:

$$\widehat{\lambda}_m^{(t+1)} = \frac{w_{\cdot\cdot}^{(t)}}{w_{\cdot m}^{(t)}}$$

$$\widehat{\theta}_m^{(t+1)} = \arg\max_{\theta} Q_m(\theta_m|\theta^{(t)})$$

EM in Exponential Families: Heuristic sidenotes

- In practice, if $f_m(y_i|\theta_m)$ does not have desirable analytical properties, then it is difficult to maximize $Q_m(\theta_m|\theta^{(t)})$.
- However, statisticians are lazy people (they put a lot of assumptions in large sample theory) thus they developed a class of parametric models with elegant properties .
- This is called **Exponential Families** (a.k.a. Koopman-Darmois-Pitman families until the late 1950s)
- Many well-known distributions belong to this family: Binomial, Beta, Gamma, Poisson, Multivariate Gaussian, Multinomial, etc.
- Such family consists most parts of classical statistics (see chapter 1 section 5, Lehmann, 1998).
- There are many equivalent definitions of exponential families, we will follow the one in Bickel and Doskum, 2001 (Section 1.6) [Because Dr. Bickel has supervised many Chinese statisticians].

EM in Exponential Families: Definition

Definition 4 (Exponential Family)

A family of distributions $\{P_\theta : \theta \in \Theta\}$, $\Theta \in \mathbb{R}^p$ is said to be a **p-parameter full rank exponential family**, if there exists real-valued functions $\eta_1, \cdots, \eta_p$ and $B(\theta)$, and functions $T_1, \cdots, T_p, h$ on $\mathbb{R}^p$ s.t. the pdf or pmf of the P_θ may be written as:

$$p(\mathbf{y}, \theta) = h(\mathbf{y}) \exp\{\sum_{j=1}^{k} \eta_j(\theta) T_j(\mathbf{y}) - B(\theta)\}, y \in \mathbb{R}^p \tag{1}$$

$$= h(\mathbf{y}) \exp\{\langle \eta(\theta), T(\mathbf{y}) \rangle - B(\theta)\} \tag{2}$$

- $\langle \cdot, \cdot \rangle$ denotes inner product in Euclidean space.
- For those who are not familiar with exponential families, I will recommend Section 2.4 in Bishop, 2006 for non-statisticians and Arash Amini's notes (start from page 62) for statisticians.

EM in Exponential Families: Properties

- Suppose $p(\mathbf{y}, \theta) = h(\mathbf{y}) \exp\{\langle \eta(\theta), T(\mathbf{y}) \rangle - B(\theta)\}$, we use another parametrization η (use η instead of θ):

$$p(\mathbf{y}, \eta) = h(\mathbf{y}) \exp\{\langle \eta, T(\mathbf{y}) \rangle - A(\eta)\}$$

Theorem 5 (Properties of exponential families)

- $\dot{A}(\eta) = \mathbb{E}(T(\mathbf{y}))$ *and* $\ddot{A}(\eta) = Var(T(\mathbf{y}))$ *where* $\dot{A}$ *means the first derivative and* $\ddot{A}$ *is the second derivative (w.r.t.* η*).*
- A *(mapping between* η *and* θ*) is convex.*
- $\eta \to \dot{A}(\eta)$ *is 1-1 on parameter space.*
- *The conjugate prior of exponential family is again exponential family.*
- *Forget about (2)(3)(4), we will only need (1).*

EM in Exponential Families: Algorithm

- **pdf for single obs**:
  $f_m(y_i|\theta_i, z_{im} = 1) = f_m(y_i|\eta_i, z_{im} = 1) = h(y_i)\exp\{\langle \eta, T(\mathbf{y})\rangle - A(\eta)\}$
- **E-step**: Recall slides 15:

$$Q_m(\eta_m|\eta^{(t)}) = \sum_{i=1}^{n} w_{im}^{(t)} [\log h(\mathbf{y_i}) + \langle \eta, T(\mathbf{y}_i)\rangle - A(\eta)] \quad (3)$$

$$= -w_{\cdot m}^{(t)} A(\eta) + \langle \eta, \sum_{i=1}^{n} w_{im}^{(t)} T(\mathbf{y}_i)\rangle + const \quad (4)$$

- **M-step**: Take derivative w.r.t. η (or θ via Chain rule), $\hat{\eta}_m^{(t+1)}$ is the solution to:

$$\sum_{i=1}^{n} w_{im}^{(t)} T(\mathbf{y}_i) = \mathbb{E}\left[\sum_{i=1}^{n} w_{im}^{(t)} T(\mathbf{y}_i)\right] = w_{\cdot m}^{(t)} \mathbb{E}[T(\mathbf{Y_1})]$$

- Since MLE is **invariant**, we have $\hat{\theta}_m^{(t+1)} = (\hat{\eta}_m^{(t+1)})^{-1}(\theta)$ (-1 denotes inverse function).

Sidenotes for EM Algorithm

- In multi-parameter cases, the "curse of dimensionality" becomes a serious issue.
- Since the numerber of parameters grows exponentially.
- PCA is often used as a first step to reduce dimensionality, but it does not help in mixture problems (Izenmann, 2013).
- This is because any class structure as exists may not be preserved by the principle components (Izenmann, 2013 and Chang, 1983).
- From my personal experience, in multivariate Gaussian case, the covariance matrix can become singular easily so EM algorithm fails.

Case Study: Classic Multivariate Gaussian Distribution

(1) Assumptions

Denote Z_i as the cluster label of y_i, this is hidden (or latent variable).

$$Z_i \sim \mathcal{M}(1, \lambda), \lambda = (\lambda_1, \cdots, \lambda_K)$$

$$Y_i | z_{im} = 1 \sim \mathcal{N}_p(\mu_m, \Sigma_m)$$

(2) Estimation

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

(3) Relationship between θ and η

For multivariate Gaussian, we have the following property:

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \eta_1 = -\frac{1}{2}\Sigma_m^{-1}, \eta_2 = \Sigma_m^{-1}\mu$$

Thus we could apply the **Invariance Principle of MLE** to derive $\hat{\mu}$ and $\widehat{\Sigma}$ via $\hat{\eta}$.

Case Study: Classic Multivariate Gaussian Distribution

Definition 6 (EM for Gaussian Mixture Model)

- **E-step**

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

- $\phi_p(\cdot)$ is the pdf of p-dimensional Gaussian distribution.
- $\widehat{W} \in [0,1]^{n \times p}$.
- Recall M-step in slide 19:

$$\mathbb{E}(T(\mathbf{Y_1})|\theta) = \mathbb{E}\begin{pmatrix} Y_1 \\ Y_1 Y_1^T \end{pmatrix} = \begin{pmatrix} \mu_m \\ \Sigma_m + \mu_m \mu_m^T \end{pmatrix}$$

Given that $Z_{1m} = 1$ (i.e. Y_1 belongs to the first cluster).

Case Study: GMM continued

Definition 7 (EM for Gaussian Mixture Model (Cont'd))

- **M-step** For $m = 1, \cdots, K$, solving:

$$\sum_i w_{im}^{(t)} = w_{\cdot m}^{(t)} \mu_m$$

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

- Iterate between E step and the following equations.

$$\mu_m^{(t+1)} = \frac{\sum_i w_{im}^{(t)} y_i}{w_{\cdot m}^{(t)}}$$

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

Prediction and simplification in GMM

- Suppose we have already derived $\hat{\lambda}, \hat{\mu}, \widehat{\Sigma}$ via EM. Note that:

$$\lambda \in \mathbb{R}^K, \mu \in \mathbb{R}^{p \times K}, \Sigma \in \mathbb{R}^{p \times p \times K}$$

- Given a new observation y, we want to predict its latent indicator z.
- $\hat{z}$ is given by:

$$\hat{z} = \arg\max_m f_m(y|\hat{\theta}, z_m = 1)$$

- **Simplification**: In practice, we often assume $\Sigma_m = \sigma_m^2 I_p$. Number of parameters goes down from $p \times p$ to 1.

Finite Mixtures as graphical models

- Now we turn into a high-level point of view (Note that Donatello Telesca is a master of graphical models (Hua Zhou, 2019)).
- Unitl 1990s, statisticians realized that many (almost all) probabilistic models can be put into a graphical model framework.
- Finite mixtures model can be viewed as either DAG(directed acyclic graph) or UG(undirected graph).
- I will demonstrate it in DAG framework.

Finite Mixtures as graphical models

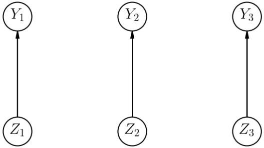


Figure: Finite Mixtures Model

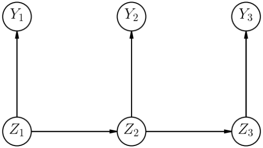


Figure: HMM and Kalman Filter

- LHS: $Z_i's$ are hidden or laten variables which generates our observed data $Y_i's$. $Z_i's$ are independent of each other.
- RHS: If we suppose Z_{i+1} is generated by Z_i, then this is called a **Hidden Markov Model** (HMM).
- If we suppose both Y and Z are continuous, then RHS is called **Kalman Filter**, which has a lot of applications in econometrics and computer vision.
- Estimation in HMM or Kalman requires more elaborate techniques known as **Baum-Welch Algorithm** (an extension of EM).
- Prediction in HMM or Kalman is called **the Viterbi Algorithm**.

K-means as a special case of GMM

Theorem 8

Assume $\Sigma_1 = \cdots = \Sigma_k = \sigma^2 I_p$. *Then as* $\sigma^2 \to 0$, *Gaussian mixture model (GMM) is equivalent to K-means clustering.*

K-means as a special case of GMM

Proof.

- Proof is based on lecture notes by Qing Zhou, 2019.
- **E-step**:

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

- As $\sigma \to 0^+$, we have:

$$w_{im}^{(t)} = \begin{cases} 1 & \text{if } m = \arg\min_j \|y_i - \mu_j^{(t)}\|_2^2 \\ 0 & \text{if else} \end{cases}$$

- **M-step**: the updated parameter is nothing but:

$$\mu_m^{(t+1)} = \frac{(\sum w_{im}^{(t)} y_i)}{|\mathcal{C}_m|}, \ \mathcal{C}_m = \{i : w_{im} = 1\}$$

END !

Next time: False Discovery Rate (empirical Bayes)
Dimension Reduction
Bayes Deconvolution and g-Modelling
Relevant Vector Machine