Expectation-Maximization (EM) Algorithm

BIOSTAT 802: Advanced Inference II

Winter, 2018

References

1. Dempster, Laird and Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. JRSSB 39, 1-38.
2. Tanner (1991). *Tools for Statistical Inference*. Lecture Notes in Statistics 67, Springer-Verlag.
3. Wu, CJF (1983). On the convergence properties of the EM algorithm. AOS 11, 95-103.

WARNING: Watch out notations, which are not completely consistent throughout the slides!

Outline

Framework

- A general approach to iterative computation of maximum-likelihood estimate when the observations can be viewed as incomplete data.
- Since each iteration of the algorithm consists of an *expectation step* followed by a *maximization step*, it is called EM algorithm.
- It is easier to present this method using some Bayesian vocabularies. Suppose observed data $Y \sim f(y|\theta)$,
  Q: Find the posterior mode $\hat{\theta}$, namely, a statistic $\hat{\theta}(y_1, \ldots, y_n)$ maximizes $f(\theta|Y)$.
- Technique: Augment the observed data Y with latent data Z so that the augmented posterior distribution $p(\theta|Y, Z)$ is “simple” in the sense that for instance, it is easy to carry out sampling/calculating/maximizing.

- Algorithm: Let $\theta^{(i)}$ be the current estimate of the mode of $p(\theta|Y)$.

  * E-step: Compute

$$
\begin{aligned}
Q(\theta, \theta^{(i)}) &= E\{\log p(\theta|Z, Y)\} \\
&\quad \text{ with respect to } p(Z|\theta^{(i)}, Y) \\
&= \int_{\mathcal{Z}} \log\{p(\theta|Z, Y)\} p(Z|\theta^{(i)}, Y) dZ.
\end{aligned}
$$

  * M-step: Maximize the Q function with respect to θ to obtain $\theta^{(i+1)}$. The algorithm is iterated until

$$
||\theta^{(i+1)} - \theta^{(i)}|| \text{ and/or } ||Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)})||
$$

  is sufficiently small.

- Explanation:

$$
\begin{aligned}
1 &= \frac{p(\theta, Z, Y)}{p(\theta, Z, Y)} = \frac{p(\theta|Z, Y)p(Z, Y)}{p(Z|\theta, Y)p(\theta|Y)p(Y)} \\
&= \frac{p(\theta|Z, Y)}{p(Z|\theta, Y)} \frac{1}{p(\theta|Y)} p(Z|Y)
\end{aligned}
$$

- Take log on both sides,

$$0 = \log p(\theta|Z, Y) - \log p(Z|\theta, Y) - \log p(\theta|Y) + \underbrace{\log p(Z|Y)}_{\text{constant with respect to } \theta.}$$

Therefore,

$$\log p(\theta|Y) = \log p(\theta|Z, Y) - \log p(Z|\theta, Y) + \text{constant}$$

Integrate both sides with respect to $p(Z|Y, \theta)$

$$\begin{aligned} \log p(\theta|Y) &= \int_{\mathcal{Z}} \log p(\theta|Z, Y) p(Z|Y, \theta) dZ - \\ &\quad \int_{\mathcal{Z}} \log p(Z|\theta, Y) p(Z|\theta, Y) dZ + \\ &\quad \int_{\mathcal{Z}} \log p(Z|Y) p(Z|\theta, Y) dZ \end{aligned}$$

where the last term is always a constant when $\theta = \theta^*$.

- Define Q function

$$Q(\theta, \theta^*) = \int_{\mathcal{Z}} \log p(\theta|Z, Y) p(Z|\theta^*, Y) dZ$$

and H function

$$\begin{aligned}
H(\theta, \theta^*) &= \int_{\mathcal{Z}} \log p(Z|\theta, Y) p(Z|\theta^*, Y) dZ \\
&= \int_{\mathcal{Z}} \log \frac{p(Z|\theta, Y)}{p(Z|\theta^*, Y)} p(Z|\theta^*, Y) dZ + \\
&\quad \int_{\mathcal{Z}} \log p(Z|\theta^*, Y) p(Z|\theta^*, Y) dZ \\
&= -KL(\theta^*, \theta) + \int_{\mathcal{Z}} \log p(Z|\theta^*, Y) p(Z|\theta^*, Y) dZ
\end{aligned}$$

where $KL(\psi, \phi) = E_\psi \log\{p(Z, \psi)/p(Z, \phi)\}$ is the Kullback-Leibler information function (or divergence function).

The Ascent Property

- Consider likelihood gain (from $\theta = \theta^{(i)}$)

$$\log\{p(\theta^{(i+1)}|Y)\} - \log\{p(\theta^{(i)}|Y)\} = Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)}) - \underbrace{(H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}))}_{\text{always } \leq 0\text{, due to Rao(1973) 1e6.6}}$$

- In fact,

$$\begin{aligned} H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}) &= KL(\theta^{(i)}, \theta^{(i)}) - KL(\theta^{(i)}, \theta^{(i+1)}) \\ &= 0 - KL(\theta^{(i)}, \theta^{(i+1)}) \\ &< 0. \end{aligned}$$

The last inequality is due to Jensen's Inequality for a strictly convex function.

- Therefore, if select $\theta^{(i+1)}$ such that $Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)})$ (exactly M-step does), then

$$p(\theta^{(i+1)}|Y) \geq p(\theta^{(i)}|Y)$$

unless

$$Q(\theta^{(i+1)}, \theta^{(i)}) = Q(\theta^{(i)}, \theta^{(i)}).$$

- It appears to be **a fixed point algorithm** that compresses the search closer to the maximum at every step.
- Where does the updating ultimately go? When will the updating stop? Under which conditions the algorithm will stop at the MLE?

Outline

Notations

Consider two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ and a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. Instead of observing the "complete data" $\boldsymbol{x} \in \mathcal{X}$, we observe the "incomplete data" $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$. Let the density function of $\boldsymbol{x}$ be $f(\boldsymbol{x}|\theta)$ with parameter $\theta \in \Theta$, and let the density of $\boldsymbol{y}$ given by

$$g(\boldsymbol{y}|\theta) = \int_{\mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\theta) d\boldsymbol{x},$$

where $\mathcal{X}(\boldsymbol{y}) = \{\boldsymbol{x} : \boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{y}\}$.

The goal is to derive the MLE of θ as $\hat{\theta} = \arg\max_{\theta \in \Theta} g(\boldsymbol{y}|\theta)$. As discussed above, in many problems, it is much simpler to maximize the complete-data specification $f(\boldsymbol{x}|\theta)$ (i.e. the M step) than the incomplete-data specification $g(\boldsymbol{y}|\theta)$ with respect to θ. And the EM algorithm provides an approach to doing so.

Since part of $\boldsymbol{x}$ is unobserved, we replace the complete-data log likelihood $\log f(\boldsymbol{x}|\theta)$ by its conditional expectation given the observed $\boldsymbol{y}$ and the current update $\theta^{(p)}$ (i.e. the E step).

The Algorithm

Let $k(\boldsymbol{x}|\boldsymbol{y},\theta) = f(\boldsymbol{x}|\theta)/g(\boldsymbol{y}|\theta)$ be conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$ and θ. Then the log-likelihood is

$$\ell(\theta') = \log g(\boldsymbol{y}|\theta') = Q(\theta'|\theta) - H(\theta'|\theta),$$

where $Q(\theta'|\theta) = E\{\log f(\boldsymbol{x}|\theta')|\boldsymbol{y},\theta\}$ and $H(\theta'|\theta) = E\{\log k(\boldsymbol{x}|\boldsymbol{y},\theta')|\boldsymbol{y},\theta\}$ are assumed to be exist for all pairs (θ,θ').

The EM algorithm proceeds $\theta^{(p)} \to \theta^{(p+1)} \in M(\theta^{(p)})$:

- E-step: Determine $Q(\theta|\theta^{(p)})$.
- M-step: Choose $\theta^{(p+1)}$ to be any value of $\theta \in \Theta$ which maximizes $Q(\theta|\theta^{(p)})$,

where $M(\theta^{(p)})$ is the set of θ values which maximizes $Q(\theta|\theta^{(p)})$ over $\theta \in \Theta$.

In other words, each iteration of the EM algorithm defines a point-to-set mapping: $\theta \to M(\theta)$ such that

$$Q(\theta'|\theta) \geq Q(\theta|\theta), \text{ for all } \theta' \in M(\theta)$$

It follows from the ascent property that

$$\ell(\theta^{(p+1)}) \geq \ell(\theta^{(p)}), \tag{1}$$

because of the inequality $H(\theta|\theta) \geq H(\theta'|\theta)$.

Where does the EM go?

According to the monotone convergence theorem, for a bounded sequence $\{\ell(\theta^{(p)})\}$, the ascent property (1) implies that $\ell(\theta^{(p)})$ converges monotonically to some ℓ^*.

Question: whether ℓ^* is the global maximum of $\ell(\theta)$ over Θ? Or, under which conditions, it may be?

First, here are assumptions required by the monotone convergence theorem:

1) Θ is a subset of the r-dimensional Euclidean space R^r;
2) $\Theta_{\theta_0} = \{\theta \in \Theta : \ell(\theta) \geq \ell(\theta_0)\}$ is compact for any $\ell(\theta_0) > -\infty$;
3) $\ell(\cdot)$ is continuous in Θ and differentiable in the interior of Θ.

Assumptions 1)-3) above imply that

$$\{\ell(\theta^{(p)})\}_{p\geq 0} \text{ is bounded above for any } \theta_0 \in \Theta$$

What makes ℓ^*?

Let $\theta^* \in \Theta$ be a value at which $\ell(\theta^*) = \ell^*$.

Question: what is the θ^*? Global maximum, local maximum or stationary point?

- There is no guarantee that θ^* is even a local maximum (thus nor the global maximum). This is because $-\nabla^2\ell(\theta^*) = -\nabla^{20}Q(\theta^*|\theta^*) + \nabla^{20}H(\theta^*|\theta^*)$, and even both $-\nabla^{20}Q$ and $-\nabla^{20}H$ are non-negative definite (n.n.d), their difference $\nabla^2\ell(\theta^*)$ is not necessarily n.n.d.
- Under some suitable conditions, θ^* may be a stationary point.

Global Convergence Theorem

Definition: A map A from points of X to subsets of X is called a *point-to-set map on* X.

Definition: A point-to-set map A is said to be *closed at* x if $x_k \to x, x_k \in X$ and $y_k \to y, y_k \in A(x_k)$ implies $y \in A(x)$.

Essentially, closeness means that either two-step updating or one-step updating ends up in the same solution set (the relay is under controlled).

Theorem (Global Convergence Theorem, (Zangwill, 1969)) Let the sequence $\{x_k\}_{k=0}^{\infty}$ be generated by $x_{k+1} \in M(x_k)$, where M is a point-to-set map on X. Let a solution set $\Gamma \subset X$ be given, and suppose that (i) all points x_k are contained in a compact set $S \subset X$; (ii) M is closed over the complement of Γ; (iii) there is a continuous function α on X such that (a) if $x \notin \Gamma, \alpha(y) > \alpha(x)$ for all $y \in M(x)$, and (b) if $x \in \Gamma, \alpha(y) \geq \alpha(x)$ for all $y \in M(x)$.
Then all the limit points of $\{x_k\}$ are in the solution set Γ and $\alpha(x_k)$ converges monotonically to $\alpha(x)$ for some $x \in \Gamma$.

Convergence of EM Algorithm

Let M be the point-to-set map in an iteration, and let $\alpha(x)$ be the log-likelihood function ℓ. The solution set Γ is

$$\begin{aligned}\mathcal{M} &= \text{set of local maxima in the interior of } \Theta\text{; or}\\ \mathcal{S} &= \text{set of stationary points in the interior of } \Theta\end{aligned}$$

Theorem 1: Let $\{\theta^{(p)}\}$ be an algorithm sequence generated by $\theta^{(p+1)} \in M(\theta^{(p)})$, and suppose that (i) M is closed point-to-set map over the complement of $\mathcal{S}$ (or $\mathcal{M}$), (ii) $\ell(\theta^{(p+1)}) > \ell(\theta^{(p)})$ for all $\theta^{(p)} \notin \mathcal{S}$ (or $\mathcal{M}$).
Then all the limit points of $\{\theta^{(p)}\}$ are stationary points (or local maxima) of ℓ, and $\ell(\theta^{(p)})$ converges monotonically to $\ell^* = \ell(\theta^*)$ for some $\theta^* \in \mathcal{S}$ (or in $\mathcal{M}$).
Remark: (i) A simple sufficient condition for the closedness of M w.r.t. $\mathcal{S}$ is that $Q(\phi|\theta)$ is continuous in both ϕ and θ. This is a very weak condition that can be satisfied in most practical situations. (ii) To establhish the closeness of M w.r.t. $\mathcal{M}$, an additional condition (eqn (11) in Wu's paper) is required.

Theorem 2 (Convergence of EM algorithm): Suppose Q satisfies the continuity condition above. Then all the limit points of $\{\theta^{(p)}\}$ are stationary points of ℓ, and $\ell(\theta^{(p)})$ converges monotonically to $\ell^* = \ell(\theta^*)$ for some θ^*.

Theorem 2 cannot be generalized to the case of local maxima because in the solution set the ascent property may hold with equality. Thus, to guarantee convergence to a local maximum, we need to impose additional conditions. Unfortunately, some strong conditions are required (Wu, 1983). One of the most popular assumptions is that the set of θ values at which ℓ^* is attained is a singleton $\{\theta^*\}$. Then $\theta^{(p)} \to \theta^*$.

Theorem 3: Suppose that $\ell(\theta)$ is unimodal in Θ with θ^* being the only stationary point and that $\nabla^{10}(\theta'|\theta)$ is continuous in θ and θ'. Then for any EM sequence $\{\theta^{(p)}\}$, $\theta^{(p)}$ converges to the unique maximizer θ^* of $\ell(\theta)$.

Remark: The singleton condition may be relaxed, to some extent, by $||\theta^{(p+1)} - \theta^{(p)}|| \to 0$ as $p \to \infty$. But this condition cannot guarantee surely $\theta^{(p)}$ converges to a local maximum, unless the solution set $\mathcal{M}$ is discrete. Thus, in the literature when using the EM algorithm, users are recommended to monitor not only $||\ell(\theta^{(p+1)}) - \ell(\theta^{(p)})|| \to 0$ but also $||\theta^{(p+1)} - \theta^{(p)}|| \to 0$.

Outline

Genetic Linkage Model (Rao, 1973)

Suppose 197 animals' genotypes are distributed into four categories as

$$Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

For example, AA, AB, BA, BB, with cell probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)$$

implicitly $\theta \in (0,1)$ is confined in $(0,1)$.

- Direct approach: using a flat prior $\theta \sim U(0,1)$. The posterior is

$$\begin{aligned} p(\theta|y_1, y_2, y_3, y_4) &= \frac{p(y_1, y_2, y_3, y_4|\theta)p(\theta)}{\int p(y_1, y_2, y_3, y_4|\theta)p(\theta)d\theta} \\ &\propto p(y_1, y_2, y_3, y_4|\theta)p(\theta) \\ &\propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}. \end{aligned}$$

Finding the posterior mode of $p(\theta|y_1, y_2, y_3, y_4)$ is equivalent to finding maximizer of the polynomial $(2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}$.

- Latent Data approach: Augment the observed data by splitting the first cell into two cells with probablities $\frac{1}{2}$ and $\frac{\theta}{4}$ respectively. The augmented data are then given by $X = (x_1, x_2, x_3, x_4, x_5)$ such that

$$x_1 + x_2 = y_1 = 125$$

$$x_{i+1} = y_i, \quad i = 2, 3, 4.$$

Also using a flat prior $\theta \sim U(0, 1)$, the posterior conditional on the augmented data is given by, through a similar augment as above,

$$\begin{aligned} p(\theta|x_1, x_2, x_3, x_4, x_5) &\propto \\ & \quad (\frac{1}{2})^{x_1} \theta^{x_2} \times \\ & \quad (1-\theta)^{x_3} (1-\theta)^{x_4} \theta^{x_5} \\ &\propto \theta^{x_2 + x_5} (1-\theta)^{x_3 + x_4}. \end{aligned}$$

By working with the augmented posterior we realize a simplification in functional form.

- EM–algorithm for this model is given as follows:
- E–step: Compute

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E \log p(\theta | Z, Y) \\ &= E\{(x_2 + x_5) \log \theta + \\ &\quad (x_3 + x_4) \log(1 - \theta) | X_2, Y\} \end{aligned}$$

where

$$\begin{aligned} p(x_2 | \theta^{(i)}, Y) &= p(x_2 | \theta^{(i)}, x_1 + x_2) \\ &\sim \text{Binomial}\left(125, \frac{\theta^{(i)}}{\theta^{(i)} + 2}\right) \end{aligned}$$

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= \{E(x_2 | \theta^{(i)}, Y) + x_5\} \log \theta \\ &\quad + (x_3 + x_4) \log(1 - \theta) \end{aligned}$$

is linear in the latent (missing) data, where

$$E(x_2 | \theta^{(i)}, Y) = 125 \frac{\theta^{(i)}}{\theta^{(i)} + 2}. \tag{2}$$

- M–step: Find $\theta^{(i+1)}$ as the solution to the following equation

$$\left.\frac{\partial Q(\theta, \theta^{(i)})}{\partial \theta}\right|_{\theta^{(i+1)}} = 0$$

$$\frac{E(X_2|\theta^{(i)}, Y) + x_5}{\theta^(i+1)} - \frac{x_3 + x_4}{1 - \theta^{(i+1)}} = 0$$

$$\theta^{(i+1)} = \frac{E(X_2|\theta^{(i)}, Y) + x_5}{E(X_2|\theta^{(i)}, Y) + x_3 + x_4 + x_5},$$

where $E(X_2|\theta(i), Y)$ is given by (2). Starting at $\theta^0 = 0.5$, that EM algorithm converges to $\theta^* = 0.6268$ (the observed posterior mode) after 4 iterations.

Outline

Direct Evaluation

Having arrived at the observed posterior mode, θ^*, one wants to evaluate the observed Fisher information given by

$$-\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2}\bigg|_{\theta=\theta^*}$$

In practice, however, this may be tedious to code or difficult to evaluate for a given data set.

Louis' Method

Due to Louis(1982)

$$
\begin{aligned}
-\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2} &= -\int_{\mathcal{Z}} \frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} p(Z|Y,\theta)dZ \\
&\quad - Var\left\{\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right\}
\end{aligned}
$$

where the variance is with respect to $p(Z|Y,\theta)$.

Monte Carlo Method

In some situation it may be difficult to analytically compute

$$\int_{\mathcal{Z}} \frac{\partial^2 \log p(\theta|Y,Z)}{\partial \theta^2} p(Z|Y,\theta) dZ$$

$$\approx \frac{1}{m} \sum_{j=1}^{m} \frac{\partial^2 \log p(\theta|Y,z_j)}{\partial \theta^2}$$

where $z_1, \ldots, z_m \overset{iid}{\sim} p(Z|\theta^*, Y)$.
Similarly, one can approximate the variance by

$$\frac{1}{m} \sum_{j=1}^{m} \left(\frac{\partial \log p(\theta|Y,z_j)}{\partial \theta} \right)^2 - \left\{ \frac{1}{m} \sum_{j=1}^{m} (\frac{\partial \log p(\theta|Y,z_j)}{\partial \theta} \right\}^2 .$$

For the example of Genetic Linkage Model,

$$\theta^* = 0.6268,\ m = 10,000,\ n = 125,\ p = \frac{\theta^*}{\theta^* + 2}.$$

The estimate variance

$$\widehat{Var}\left(\frac{\partial \log p(\theta | Y, Z)}{\partial \theta}\right) = 57.8.$$

Outline

Derivations

- Consider a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

  where $\epsilon \sim N(0, \sigma^2)$.
- Data observed are pairs:

$$(y_i, \boldsymbol{x}_i), i = 1, \ldots, n.$$

- Missing data can arise from either the response or covariates. Let us consider the case of missing in response. So, write

$$\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}) \sim N_n(\boldsymbol{\mu}, \sigma^2 I),$$

  where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. Let $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$.
- Parameters to be estimated are $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and σ^2.

- For the ease of exposition, re-arrange the responses as

$$\boldsymbol{y} = (\underbrace{y_1, \ldots, y_{m_0}}_{\text{missing}}, \underbrace{y_{m_0+1}, \ldots, y_n}_{\text{observed}})^T.$$

Clearly, for the normal distribution,

$$S(\boldsymbol{y}) = (y_i, i = 1, \ldots, n; y_i^2, i = 1, \ldots, n)$$

gives a set of sufficient statistics.

- E-Step: Calculate conditional expectations of sufficient statistics:

$$y_i^{(r)} = E(y_i | \boldsymbol{y}_{obs}, \boldsymbol{X}, \beta^{(r)}, \sigma^{2(r)})$$
$$\begin{cases} y_i, & \text{if } y_i, i = m_0 + 1, \ldots, n \text{ observed} \\ \boldsymbol{x}_i^T \beta^{(r)}, & \text{if } y_i, i = 1, \ldots, m_0 \text{ missing} \end{cases}$$

And

$$y_i^{2(r)} = E(y_i^2 | \boldsymbol{y}_{obs}, \boldsymbol{X}, \beta^{(r)}, \sigma^{2(r)})$$
$$\begin{cases} y_i^2 & \text{if } y_i \text{ observed} \\ \sigma^{2(r)} + \{\boldsymbol{x}_i^T \beta^{(r)}\}^2 & \text{if } y_i \text{ missing} \end{cases}$$

- M-step: Find the MLE based on the full data $\boldsymbol{y}^{(r)}$ and $\boldsymbol{y}^{2(r)}$.

$$\beta^{(r+1)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}^{(r)}$$

- Note that this update for $\beta^{(r+1)}$ doesn't involve $\sigma^{2(r+1)}$, so one can update $\sigma^{2(r+1)}$ at the very end when the update of $\beta^{(r+1)}$ is complete.
- Update $\sigma^{2(r+1)}$ by

$$\sigma^{2(r+1)} = \frac{1}{n}\left\{m_0\sigma^{2(r)} + \sum_{i=m_0+1}^{n}(y_i - \boldsymbol{x}_i^T\beta^{(r)})^2\right\}$$

- At convergence, $\beta^{(*)}$ is obtained, and then plugged in

$$\sigma^{2(*)} = \frac{1}{n}\left\{m_0\sigma^{2(*)} + \sum_{i=m_0+1}^{n}(y_i - \boldsymbol{x}_i^T\beta^{(*)})^2\right\}$$

Solving for $\sigma^{2(*)}$ leads to

$$\sigma^{2(*)} = \frac{1}{n-m_0}\sum_{i=m_0+1}^{n}(y_i - \boldsymbol{x}_i^T\beta^{(*)})^2.$$

Outline

Framework

The density function of the normal-normal mixture takes the following form:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \left[p_1 \exp\left\{ -\frac{(x_i - \mu_1)^2}{2\sigma^2} \right\} + p_2 \exp\left\{ -\frac{(x_i - \mu_2)^2}{2\sigma^2} \right\} \right],$$

where p_k is the probability that model $N(\mu_k, \sigma^2)$ is observed, $k = 1, 2$, so $p_1 + p_2 = 1$.
We want to derive the maximum likelihood estimation for parameters, $p_1, p_2 = 1 - p_1, \mu_1, \mu_2, \sigma^2$, which will be denoted by θ.

Likelihood Augmentation by Latent Variable

Define a latent varaible that indicates the choice of a normal model as follows:

$$Z_i = \begin{cases} 1, & \text{if } X_i \sim N(\mu_1, \sigma^2); \\ 0, & \text{if } X_i \sim N(\mu_2, \sigma^2) \end{cases}$$

which is obviously a Bernoulli random variable with $p_1 = P(Z_i = 1), p_2 = P(Z_i = 0)$. Consequently we can write the conditional density as

$$p(x_i|z_i) = \{\phi(x_i; \mu_1, \sigma)\}^{z_i}\{\phi(x_i; \mu_2, \sigma)\}^{1-z_i}, \ z_i = 0, 1.$$

Then the augmented likelihood is given by

$$\begin{aligned} p(\theta|x_i, z_i, i = 1, \ldots n) &= \prod_{i=1}^{n} f(x_i|z_i)f(z_i) \\ &= \prod_{i=1}^{n} \{p_1\phi(x_i; \mu_1, \sigma)\}^{z_i} \{p_2\phi(x_i; \mu_2, \sigma)\}^{1-z_i}. \end{aligned}$$

The posterior of the latent variable on the observed data and parameters are:

$$
\begin{aligned}
P(Z_i = 1|\theta, x_i) &= \frac{P(Z_i = 1)f(x_i|y_i = 1, \theta)}{P(Z_i = 0)f(x_i|z_i = 0, \theta) + P(Z_i = 1)f(x_i|z_i = 1, \theta)} \\
&= \frac{p_1\phi(x_i; \mu_1, \sigma)}{p_1\phi(x_i; \mu_1, \sigma) + p_2\phi(x_i; \mu_2, \sigma)} \\
&\overset{def}{=} \pi_1(x_i; \theta) \\
P(Z_i = 0|\theta, x_i) &= \frac{p_2\phi(x_i; \mu_2, \sigma)}{p_1\phi(x_i; \mu_1, \sigma) + p_2\phi(x_i; \mu_2, \sigma)} \\
&\overset{def}{=} \pi_2(x_i; \theta) = 1 - \pi_1(x_i; \theta).
\end{aligned}
$$

We can rewrite this as

$$
f(z_i|x_i,\theta) = \pi_1(x_i; \theta)^{z_i}\pi_2(x_i; \theta)^{1-z_i}.
$$

Then given the updated value at iteration j available, we have $\theta^{(j)} = (p_1^{(j)}, p_2^{(j)}, \mu_1^{(j)}, \mu_2^{(j)}, \sigma^{(j)})$ and $\pi(x_i, \theta^{(j)})$.

Derivation: E-Step

Then the Q-function is given by

$$
\begin{aligned}
Q(\theta, \theta^{(j)}) &= \sum_{i=1}^{n} \sum_{z_i \in \{0,1\}} \log p(x_i, z_i; \theta) p(z_i | x_i, \theta^{(j)}) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{2} \frac{p_k^{(j)} \phi(x_i; \mu_k^{(j)}, \sigma^{(j)})}{p_1^{(j)} \phi(x_i; \mu_1^{(j)}, \sigma^{(j)}) + p_2^{(j)} \phi(x_i; \mu_2^{(j)}, \sigma^{(j)})} \log \left(p_k \phi(x_i; \mu_k, \sigma) \right) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{2} \pi_k(x_i, \theta^{(j)}) \log \left(p_k \phi(x_i; \mu_k, \sigma) \right)
\end{aligned}
$$

In the E step, we evaluate both $\pi_k(x_i, \theta^{(j)})$ and $Q(\theta, \theta^{(j)})$.

Derivation: M-Step

In the M step, we find $\theta^{(j+1)} = \arg\max_{\theta} Q(\theta, \theta^{(j)})$ by taking the derivative with respect to p_1, μ_1, μ_2 and σ and setting to 0:

$$\frac{\partial}{\partial p_1} Q(\theta, \theta^{(j)}) = \sum_{i=1}^{n} \pi_1(x_i, \theta^{(j)}) \frac{1}{p_1} - \sum_{i=1}^{n} \pi_2(x_i, \theta^{(j)}) \frac{1}{1 - p_1} = 0,$$

$$\frac{\partial}{\partial \mu_k} Q(\theta, \theta^{(j)}) = \sum_{i=1}^{n} \pi_k(x_i, \theta^{(j)}) \frac{\frac{\partial}{\partial \mu_k} \phi(x_i; \mu_k, \sigma)}{\phi(x_i; \mu_k, \sigma)} = 0$$

$$\frac{\partial}{\partial \sigma} Q(\theta, \theta^{(j)}) = \sum_{i=1}^{n} \sum_{k=1}^{2} \pi_k(x_i, \theta^{(j)}) \frac{\frac{\partial}{\partial \sigma} \phi(x_i; \mu_k, \sigma)}{\phi(x_i; \mu_k, \sigma)} = 0$$

The closed form expressions of the solution to the above equations are

$$p_k^{(j+1)} = \frac{\sum_{i=1}^n \pi_k(x_i, \theta^{(j)})}{n}, k = 1, 2;$$

$$\mu_k^{(j+1)}) = \frac{\sum_{i=1}^n \pi_k(x_i, \theta^{(j)})x_i}{\sum_{i=1}^n \pi_k(x_i, \theta^{(j)})}, k = 1, 2;$$

$$\sigma^{2(j+1)} = \frac{\sum_{i=1}^n \sum_{k=1}^2 \pi_k(x_i, \theta^{(j)})(x_i - \mu_k^{(j+1)})^2}{n}.$$

Initial Values

To specify the initial values, we may first run a two-class clustering analysis, from which we can estimate $p_k^{(0)}$ = the proportion of data points classified into class k, and $\mu_k^{(0)}$ = the class-specific sample mean, $k = 1, 2$, and $\sigma^{(0)}$ = the sample variance of the overall data.