Statistical Models & Computing Methods

Lecture 7: Expectation Maximization

Cheng Zhang

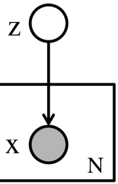
School of Mathematical Sciences, Peking University

November 19, 2020

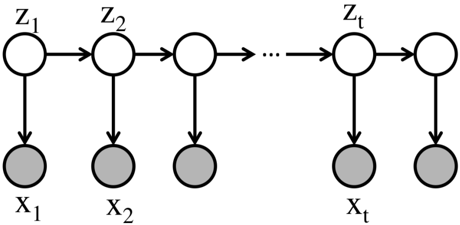
- In this lecture, we discuss Expectation-Maximization (EM), which is an iterative optimization method dealing with missing or latent data.
- In such cases, we may assume the observed data x are generated from random variable X along with missing or unobserved data z from random variable Z. We envision complete data would have been $y = (x, z)$.
- Very often, the inclusion of the observed data z is a *data augmentation* strategy to ease computation. In this case, Z is often referred to as *latent* variable.

- Some of the variables in the model are not observed.
- Examples: mixture model, hidden Markov model (HMM), latent Dirichlet allocation (LDA), etc.
- We consider the learning problem of latent variable models

Mixture Model



Hidden Markov Model



Marginal Likelihood

- complete data likelihood $p(x, z|\theta)$, θ is model parameter
- When z is missing, we need to marginalize out z and use the marginal log-likelihood for learning

$$\log p(x|\theta) = \log \sum_z p(x, z|\theta)$$

- Examples: Gaussian mixture model. $z \sim \mathrm{Discrete}(\pi)$, $\theta = (\pi, \mu, \Sigma)$

$$\begin{aligned} p(x|\theta) &= \sum_k p(z = k|\theta) p(x|z = k, \theta) \\ &= \sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \\ &= \sum_k \pi_k \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right) \end{aligned}$$

Learning in Latent Variable Model

- For most of these latent variable models, when the missing components z are observed, the complete data likelihood often factorizes, and the maximum likelihood estimates hence have closed-form solutions.
- When z are not observed, marginalization destroys the factorizible structure and makes learning much more difficult.
- How to learn in this scenario?
  - **Idea 1**: simply take derivative and use gradient ascent directly
  - **Idea 2**: find appropriate estimates of z (e.g., using the current conditional distribution $p(z|x, \theta)$), fill them in and do complete data learning – This is EM!

Expectation Maximization

- At each iteration, the EM algorithm involves two steps
  - based on the current $\theta^{(t)}$, fill in unobserved z to get *complete data* (x, z')
  - Update θ to maximize the complete data log-likelihood $\ell(x, z'|\theta) = \log p(x, z'|\theta)$
- How to choose z'?
  - Use conditional distribution $p(z|x, \theta^{(t)})$
  - Take full advantage of the current estimates $\theta^{(t)}$

$$\mathbb{E}_{p(z|x,\theta^{(t)})}\ell(x, z|\theta) = \sum_z p(z|x, \theta^{(t)})\ell(x, z|\theta)$$

In some sense, this is our best guess (as shown later).

More specifically, we start from some initial $\theta^{(0)}$. In each iteration, we follow the two steps below

- **Expectation (E-step)**: compute $p(z|x, \theta^{(t)})$ and form the expectation using the current estimate $\theta^{(t)}$

$$Q^{(t)}(\theta) = \mathbb{E}_{p(z|x,\theta^{(t)})} \ell(x, z|\theta)$$

- **Maximization (M-step)**: Find θ that maximizes the expected complete data log-likelihood

$$\theta^{(t+1)} = \arg\max_{\theta} Q^{(t)}(\theta)$$

In many cases, the expectation is easier to handle than the marginal log-likelihood.

How does EM Work?

- EM algorithm can be viewed as optimizing a lower bound on the marginal log-likelihood $\mathcal{L}(\theta) = \log p(x|\theta)$
- A class of lower bounds

$$
\begin{aligned}
\mathcal{L}(\theta) = \log \sum_z p(x, z|\theta) &= \log \sum_z q(z) \frac{p(x, z|\theta)}{q(z)} \\
&\geq \sum_z q(z) \log \frac{p(x, z|\theta)}{q(z)} \qquad \text{- Jensen's inequality} \\
&= \sum_z q(z) \log p(x, z|\theta) - \sum_z q(z) \log q(z), \quad \forall q(z)
\end{aligned}
$$

- The term in the last equation is often called *Free-energy*

$$\mathcal{F}(q, \theta) = \sum_z q(z) \log p(x, z|\theta) - \sum_z q(z) \log q(z)$$

Lower Bound Maximization

- Free-energy is a lower bound of the true log-likelihood

$$\mathcal{L}(\theta) \geq \mathcal{F}(q, \theta)$$

- EM is simply doing coordinate ascent on $\mathcal{F}(q, \theta)$
  - E-step: Find $q^{(t)}$ that maximizes $\mathcal{F}(q, \theta^{(t)})$
  - M-step: Find $\theta^{(t+1)}$ that maximizes $\mathcal{F}(q^{(t)}, \theta)$
- Properties:
  - Each iteration improves $\mathcal{F}$

$$\mathcal{F}(q^{(t+1)}, \theta^{(t+1)}) \geq \mathcal{F}(q^{(t)}, \theta^{(t)})$$

  - Each iteration improves $\mathcal{L}$ as well

$$\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{L}(\theta^{(t)})$$

    will show later

- Find q that maximizes $\mathcal{F}(q, \theta^{(t)})$

$$
\begin{aligned}
\mathcal{F}(q, \theta) &= \sum_z q(z) \log p(x, z|\theta) - \sum_z q(z) \log q(z) \\
&= \sum_z q(z) \log \frac{p(z|x, \theta) p(x|\theta)}{q(z)} \\
&= \sum_z q(z) \log \frac{p(z|x, \theta)}{q(z)} + \log p(x|\theta) \\
&= \mathcal{L}(\theta) - D_{\mathrm{KL}}\left(q(z) \| p(z|x, \theta)\right) \\
&\leq \mathcal{L}(\theta)
\end{aligned}
$$

$$\mathcal{F}(q,\theta^{(t)}) = \mathcal{L}(\theta^{(t)}) - D_{\mathrm{KL}}(q(z)\|p(z|x,\theta^{(t)}))$$

- KL divergence is non-negative and is minimized (equals to 0) iff the two distributions are identical.
- Therefore, $\mathcal{F}(q,\theta^{(t)})$ is maximized at $q^{(t)}(z) = p(z|x,\theta^{(t)})$.
- So when we are computing $p(z|x,\theta^{(t)})$, we are actually computing $\arg\max_q \mathcal{F}(q,\theta^{(t)})$
- Moreover,

$$\mathcal{F}(q^{(t)},\theta^{(t)}) = \mathcal{L}(\theta^{(t)})$$

  this means the lower bound matches the true log-likelihood at $\theta^{(t)}$, which is crucial for the improvement on $\mathcal{L}$.

M-step

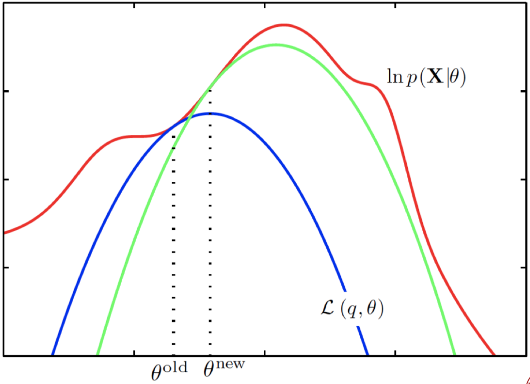
- Find $\theta^{(t+1)}$ that maximizes $\mathcal{F}(q^{(t)}, \theta)$

$$
\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \mathcal{F}(q^{(t)}, \theta) \\
&= \arg\max_{\theta} \sum_{z} p(z|x, \theta^{(t)}) \log p(x, z|\theta) + H(p(z|x, \theta^{(t)})) \\
&= \arg\max_{\theta} \mathbb{E}_{p(z|x,\theta^{(t)})} \ell(x, z|\theta)
\end{aligned}
$$

- The expected complete data log-likelihood usually can be solved in the same manner (closed-form solutions) as the fully-observed model.

Monotonicity of EM

$$
\begin{aligned}
\mathcal{L}(\theta^{(t+1)}) &\geq \mathcal{F}(q^{(t)}, \theta^{(t+1)}) \\
&\geq \mathcal{F}(q^{(t)}, \theta^{(t)}) = \mathcal{L}(\theta^{(t)})
\end{aligned}
$$



EM for Exponential Families

- When the complete data follow an exponential family distribution (in canonical form), the density is

$$p(x, z|\theta) = h(x, z) \exp(\theta \cdot T(x, z) - A(\theta))$$

- E-step

$$\begin{aligned} Q^{(t)}(\theta) &= \mathbb{E}_{p(z|x,\theta^{(t)})} \log p(x, z|\theta) \\ &= \theta \cdot \mathbb{E}_{p(z|x,\theta^{(t)})} T(x, z) - A(\theta) + \text{Const} \end{aligned}$$

- M-step

$$\nabla_\theta Q^{(t)}(\theta) = 0 \Rightarrow \mathbb{E}_{p(z|x,\theta^{(t)})} T(x, z) = \nabla_\theta A(\theta) = \mathbb{E}_{p(x,z|\theta)} T(x, z)$$

Examples: Censored Survival Times

- In survival analyses, we often have to terminate our study before observing the real survival times, leading to censored survival data.
- Suppose the observed data are $Y = \{(t_1, \delta_1), \ldots, (t_n, \delta_n)\}$, where $T_j \sim \mathrm{Exp}(\mu)$ and δ_j is the indicator of a censored sample. WLOG, assume $\delta_i = 0, i \leq r, \quad \delta_i = 1, i > r$
- The log-likelihood function is

$$\begin{aligned}\log p(Y|\mu) &= \sum_{i=1}^{r} \log p(t_i|\mu) + \sum_{i>r} \log p(T_i > t_i|\mu) \\ &= -r \log \mu - \sum_{i=1}^{n} t_i/\mu\end{aligned}$$

- The MLE of μ: $\hat{\mu} = \sum_{i=1}^{n} t_i / r$

- Let us see how EM works in this simple case.
- Let $t = (T_1, \dots, T_n) = (T_1, \dots, T_r, z)$ be the complete data vector, where $z = (T_{r+1}, \dots, T_n)$ are the unobserved $n - r$ censored random variables.
- Natural parameter $1/\mu$, sufficient statistics $\sum_{i=1}^n T_i$, and $\mathbb{E}_\mu \sum_{i=1}^n T_i = n\mu$
- By the lack of memory, $T_i | T_i > t_i \sim t_i + \mathrm{Exp}(\mu)$, $\forall i > r$.

$$\mathbb{E}_{p(z|Y,\mu^{(k)})} \sum_{i=1}^n T_i = \sum_{i=1}^r t_i + \sum_{i>r} t_i + (n-r)\mu^{(k)}$$

- Update formula

$$\mu^{(k+1)} = \frac{\sum_{i=1}^n t_i + (n-r)\mu^{(k)}}{n}$$

Gaussian Mixture Model

- Consider clustering of data $X = \{x_1, \dots, x_N\}$ using a finite mixture of Gaussians.

$$z \sim \text{Discrete}(\pi), \quad x|z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

$\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ are model parameters

- Complete data log-likelihood

$$\begin{aligned}\log p(x, z|\theta) &= \log \prod_{k=1}^{K} \left(p(z = k)p(x|z = k)\right)^{1_{z=k}} \\ &= \sum_{k=1}^{K} 1_{z=k}(\log \pi_k + \log \mathcal{N}(x|\mu_k, \Sigma_k))\end{aligned}$$

E-step

- Compute the conditional probability $p(z_n|x_n, \theta^{(t)})$ via Bayes' theorem

$$p(z_n|x_n, \theta) = \frac{p(z_n, x_n|\theta)}{\sum_{z_n} p(z_n, x_n|\theta)}$$

$$p(z_n = k|x_n, \theta^{(t)}) = \frac{\pi_k^{(t)} \mathcal{N}(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_k \pi_k^{(t)} \mathcal{N}(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})}$$

- Denote $\gamma_{n,k}^{(t)} \triangleq p(z_n = k|x_n, \theta^{(t)})$, which can be viewed as a *soft clustering* of x_n

$$\sum_k \gamma_{n,k}^{(t)} = 1$$

- Expected complete-data log-likelihood

$$
\begin{aligned}
Q^{(t)}(\theta) &= \sum_n \sum_{z_n} p(z_n|x_n, \theta^{(t)}) \log p(x_n, z_n|\theta) \\
&= \sum_n \sum_k \gamma_{n,k}^{(t)} \left(\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k)\right) \\
&= \sum_k \sum_n \gamma_{n,k}^{(t)} \left(\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)
\end{aligned}
$$

Substitute $\mathcal{N}(x_n|\mu_k, \Sigma_k)$ in

$$
\begin{aligned}
Q^{(t)}(\theta) = \sum_k \sum_n \gamma_{n,k}^{(t)} \Big(& \log \pi_k - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| \\
& - \frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k)\Big)
\end{aligned}
$$

M-step

- Maximize $Q^{(t)}(\theta)$ with respect to π using Lagrange multipliers

$$\pi_k^{(t+1)} \propto \sum_n \gamma_{n,k}^{(t)}$$

  Therefore

$$\pi_k^{(t+1)} = \frac{\sum_n \gamma_{n,k}^{(t)}}{\sum_k \sum_n \gamma_{n,k}^{(t)}} = \frac{\sum_n \gamma_{n,k}^{(t)}}{\sum_n \sum_k \gamma_{n,k}^{(t)}} = \frac{\sum_n \gamma_{n,k}^{(t)}}{N}$$

- Note that $\sum_n \gamma_{n,k}^{(t)}$ can be viewed as the weighted number of data points in mixture component k, and $\pi_k^{(t+1)}$ is the fraction of data the belongs to mixture component k.

- Compute the derivative w.r.t μ_k

$$\frac{\partial Q^{(t)}(\theta)}{\partial \mu_k} = \sum_n \gamma_{n,k}^{(t)} \Sigma_k^{-1}(x_n - \mu_k) = \Sigma_k^{-1} \sum_n \gamma_{n,k}^{(t)}(x_n - \mu_k)$$

- Therefore,

$$\mu_k^{(t+1)} = \frac{\sum_n \gamma_{n,k}^{(t)} x_n}{\sum_n \gamma_{n,k}^{(t)}}$$

$\mu_k^{(t+1)}$ is the weighted mean of data points assigned to mixture component k

- Similarly, we can get

$$\Sigma_k^{(t+1)} = \frac{\sum_n \gamma_{n,k}^{(t)}(x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_n \gamma_{n,k}^{(t)}}$$

- **E-step**: Compute the soft clustering probabilities

$$\gamma_{n,k}^{(t)} = \frac{\pi_k^{(t)} \mathcal{N}(x_n | \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_k \pi_k^{(t)} \mathcal{N}(x_n | \mu_k^{(t)}, \Sigma_k^{(t)})}$$

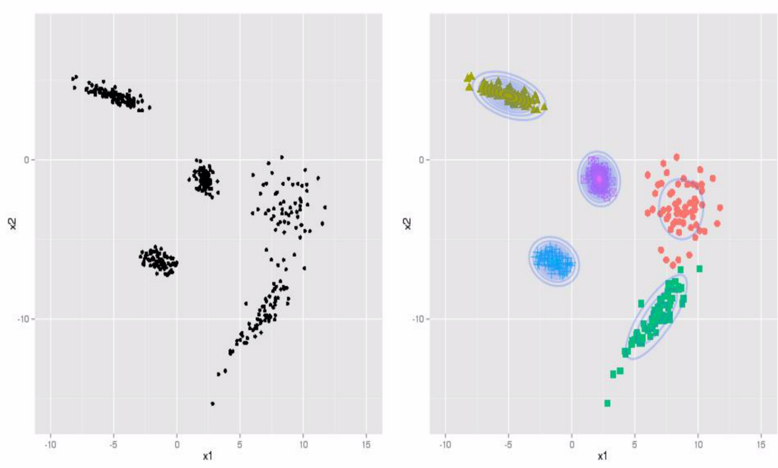
- **M-step**: Update parameters

$$\pi_k^{(t+1)} = \frac{\sum_n \gamma_{n,k}^{(t)}}{N}$$

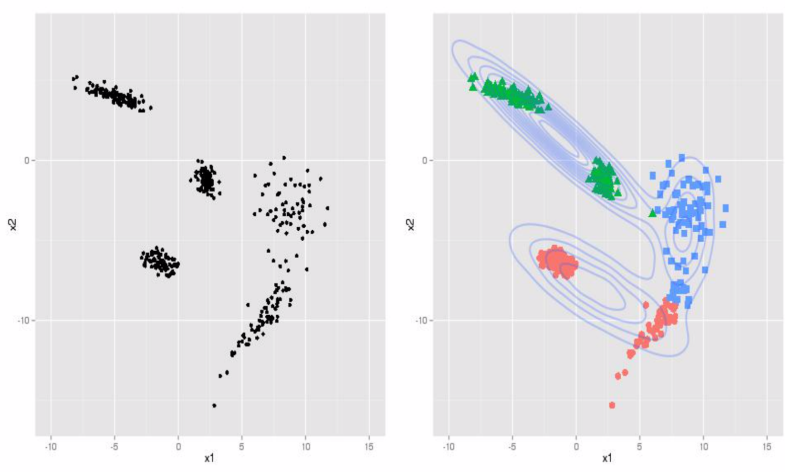
$$\mu_k^{(t+1)} = \frac{\sum_n \gamma_{n,k}^{(t)} x_n}{\sum_n \gamma_{n,k}^{(t)}}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_n \gamma_{n,k}^{(t)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_n \gamma_{n,k}^{(t)}}$$

Examples: Mixture of 5 Gaussians



Examples: Mixture of 3 Gaussians



- The k-means algorithm follows two steps
  - Assignment step: assign data to the nearest cluster

$$\gamma_{n,k} = \begin{cases} 1, & k = \arg\min_{k'} \|x_n - \mu_{k'}\| \\ 0, & \text{otherwise} \end{cases}$$

  - Update step: set μ_k to the mean of data points assigned to the k-th cluster

$$\mu_k = \frac{\sum_n \gamma_{n,k}^{(t)} x_n}{\sum_n \gamma_{n,k}^{(t)}} = \frac{1}{N_k} \sum_{n:\gamma_{n,k}=1} x_n$$

  N_k is the number of data points assigned to the k-th cluster.

- Therefore, k-means can be viewed as a special case of EM for Gaussian mixture models where $\Sigma_k = I$ and $\gamma_{n,k}$ are hard assignments instead of soft clustering probabilities.

- Sequence data $x_1, x_2, \ldots, x_T$, each $x_n \in \mathbb{R}^d$
- Hidden variables $z_1, z_2, \ldots, z_T$, each $z_t \in \{1, 2, \ldots, K\}$
- Joint probability

$$p(x, z) = p(z_1) \prod_{t=1}^{T-1} p(z_{t+1}|z_t) \prod_{t=1}^{T} p(x_t|z_t)$$

- $p(x_t|z_t)$ is the *emission probability*, could be a Gaussian

$$p(x_t|z_t = k) = \mathcal{N}(x_t|\mu_k, \Sigma_k)$$

- $p(z_{t+1}|z_t)$ is the *transition probability*, a $K \times K$ matrix
  $a_{ij} = p(z_{t+1} = j|z_t = i)$, $\sum_j a_{ij} = 1$
- $p(z_1) \sim \text{Discrete}(\pi)$ is the prior for the first hidden state

Expected Complete Data Log-likelihood

- The expected complete data log-likelihood is

$$
\begin{aligned}
Q &= \mathbb{E}_{p(z|x)} \log p(x,z) \\
&= \sum_{z} p(z|x) \left(\log p(z_1) + \sum_{t=1}^{T-1} \log p(z_{t+1}|z_t) + \sum_{t=1}^{T} \log p(x_t|z_t) \right) \\
&= \sum_{z_1} p(z_1|x) \log p(z_1) + \sum_{t=1}^{T-1} \sum_{z_t, z_{t+1}} p(z_t, z_{t+1}|x) \log p(z_{t+1}|z_t) \\
&\qquad + \sum_{t=1}^{T} \sum_{z_t} p(z_t|x) \log p(x_t|z_t)
\end{aligned}
$$

- Therefore, in the E-step, we need to compute unary and pairwise marginal probabilities $p(z_t|x)$ and $p(z_t, z_{t+1}|x)$.

E-step: Forward-Backward Algorithm

- Using the sequential structure of HMM, we can compute these marginal probabilities via **dynamic programming**.
- The **forward algorithm**

$$
\begin{aligned}
\alpha_{t+1}(j) &= p(z_{t+1} = j, x_1, \ldots, x_{t+1}) \\
&= \sum_i p(z_{t+1} = j, z_t = i, x_1, \ldots, x_{t+1}) \\
&= p(x_{t+1}|z_{t+1} = j) \sum_i p(z_{t+1} = j|z_t = i) p(z_t, x_1, \ldots, x_t) \\
&= p(x_{t+1}|z_{t+1} = j) \sum_i a_{ij} p(z_t, x_1, \ldots, x_t) \\
&= p(x_{t+1}|z_{t+1} = j) \sum_i a_{ij} \alpha_t(i)
\end{aligned}
$$

E-step: Forward-Backward Algorithm

- The **backward algorithm**

$$\begin{aligned}\beta_t(i) &= p(x_{t+1}, \ldots, x_T | z_t = i) \\ &= \sum\nolimits_j p(x_{t+1}, \ldots, x_T, z_{t+1} = j | z_t = i) \\ &= \sum\nolimits_j a_{ij} p(x_{t+1} | z_{t+1} = j) \beta_{t+1}(j)\end{aligned}$$

- Unary marginal probability

$$p(z_t = j | x) \propto p(z_t = j, x) = \alpha_t(j) \beta_t(j)$$

- Pairwise marginal probability

$$\begin{aligned}p(z_{t+1} = j, z_t = i | x) \propto p(z_{t+1} = j, z_t = i, x) \\ = \alpha_t(i) a_{ij} p(x_{t+1} | z_{t+1} = j) \beta_{t+1}(j)\end{aligned}$$

- From the E-step, we have

$$\gamma_{t,k} = p(z_t = k|x) = \frac{\alpha_t(k)\beta_t(k)}{\sum_k \alpha_t(k)\beta_t(k)}$$

$$\xi_t(i,j) = p(z_{t+1} = j, z_t = i|x) = \frac{\alpha_t(i)a_{ij}p(x_{t+1}|z_{t+1} = j)\beta_{t+1}(j)}{\sum_k \alpha_t(k)\beta_t(k)}$$

- The expected complete data log-likelihood is

$$Q = \sum_k \gamma_{1,k} \log \pi_k + \sum_{t=1}^{T-1} \sum_{i,j} \xi_t(i,j) \log a_{ij} + \sum_{t=1}^{T} \sum_k \gamma_{t,k} \log \mathcal{N}(x_t|\mu_k, \Sigma_k)$$

- Closed form solution for M-step – just like in the Gaussian mixture model

EM algorithm finds MLE for models with missing/latent variables. Applicable if the following pieces are easy to solve

- Estimating missing data from observed data using current parameters (E-step)
- Find complete data MLE (M-step)

Pros

- No need for gradients, learning rates, etc.
- Fast convergence
- Monotonicity. Guaranteed to improve $\mathcal{L}$ at every iteration

Cons

- Can get stuck at local optimal
- Requires conditional distribution $p(z|x, \theta)$ to be tractable

- While EM increases the marginal likelihood in each iteration and often converges to a stationary point, we are not clear about the convergence rate and how does that relate to the missing data scenario.
- Moreover, the requirements of tractable conditional distribution and easy complete data MLE may be too restrictive in practice.
- In what follows, we will discuss the convergence theory for EM and introduce some variants of it that can be applied in more general settings.

Example: Censored Survival Times

- Recall that in the censored survival times example, given the observed data $Y = \{(t_1, \delta_1), \ldots, (t_n, \delta_n)\}$, where t_j follows an exponential distribution with mean μ and can be either censored or not as indicated by δ_j.
- Assume $\delta_i = 0, i \leq r,\ \delta_i = 1, i > r$. The MLE of μ is $\hat{\mu} = \sum_{i=1}^{n} t_i / r$
- EM update formula

$$\mu^{(k+1)} = \frac{\sum_{i=1}^{n} t_i + (n-r)\mu^{(k)}}{n}$$

- Therefore,

$$\mu^{(k+1)} - \hat{\mu} = \frac{n-r}{n}(\mu^{(k)} - \hat{\mu})$$

  Linear convergence, rate depends on the amount of missing information

EM as A Fixed Point Algorithm

We can view EM update as a map

$$\theta^{(t+1)} = \Phi(\theta^{(t)}), \quad \Phi(\theta) = \arg\max_{\theta'} Q(\theta'|\theta)$$

where $Q(\theta'|\theta) = \mathbb{E}_{p(z|x,\theta)} \log p(x, z|\theta')$

Lemma 1

If for some θ^*, $\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta)$, $\forall\theta$, then for every EM algorithm

$$\mathcal{L}(\Phi(\theta^*)) = \mathcal{L}(\theta^*),\ Q(\Phi(\theta^*)|\theta^*) = Q(\theta^*|\theta^*)$$

and

$$p(z|x, \Phi(\theta^*)) = p(z|x, \theta^*),\ \text{a.s.}$$

Lemma 2

If for some θ^*, $\mathcal{L}(\theta^*) > \mathcal{L}(\theta)$, $\forall\theta \neq \theta^*$, then for every EM algorithm

$$\Phi(\theta^*) = \theta^*$$

Theorem 1

Suppose that $\theta^{(t)}, t = 0, 1, \ldots$ is an instance of an EM algorithm such that

- the sequence $\mathcal{L}(\theta^{(t)})$ is bounded
- for some $\lambda > 0$ and all t,

$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) \geq \lambda(\theta^{(t+1)} - \theta^{(t)})(\theta^{(t+1)} - \theta^{(t)})^T$$

Then the sequence $\theta^{(t)}$ converges to some θ^*

- Since $\theta^{(t+1)} = \Phi(\theta^{(t)})$ maximizes $Q(\theta'|\theta^{(t)})$, we have

$$\frac{\partial Q}{\partial \theta'}(\theta^{(t+1)}|\theta^{(t)}) = 0$$

- For all t, there exists a $0 \le \alpha_0^{(t+1)} \le 1$ such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) = -(\theta^{(t+1)} - \theta^{(t)}) \cdot \frac{\partial^2 Q}{\partial \theta'^2}(\theta_0^{(t+1)}|\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)})^T$$

  where $\theta_0^{(t+1)} = \alpha_0 \theta^{(t)} + (1 - \alpha_0)\theta^{(t+1)}$
- If the sequence $\frac{\partial^2 Q}{\partial \theta'^2}(\theta_0^{(t+1)}|\theta^{(t)})$ is negative definite with eigenvalues bounded away from zero and $\mathcal{L}(\theta^{(t)})$ is bounded, by Theorem 1, $\theta^{(t)}$ converges to some θ^*

- When EM converges, it converges to a fixed point of the map

$$\theta^* = \Phi(\theta^*)$$

- Taylor expansion of Φ at θ^* yields

$$\theta^{(t+1)} - \theta^* = \Phi(\theta^{(t)}) - \Phi(\theta^*) \approx \nabla\Phi(\theta^*)(\theta^{(t)} - \theta^*)$$

- The global rate of EM defined as

$$\rho = \lim_{t\to\infty} \frac{\|\theta^{(t+1)} - \theta^*\|}{\|\theta^{(t)} - \theta^*\|}$$

  equals the largest eigenvalue of $\nabla\Phi(\theta^*)$ and $\rho < 1$ when the observed Fisher information $-\nabla^2\mathcal{L}(\theta^*)$ is positive definite.

Proof

- As aforementioned, $\Phi(\theta)$ maximize $Q(\theta'|\theta)$, therefore

$$\frac{\partial Q}{\partial \theta'}(\Phi(\theta)|\theta) = 0, \quad \forall \theta$$

- Differentiate w.r.t. θ

$$\frac{\partial^2 Q}{\partial \theta'^2}(\Phi(\theta)|\theta)\nabla\Phi(\theta) + \frac{\partial^2 Q}{\partial \theta \partial \theta'}(\Phi(\theta)|\theta) = 0$$

let $\theta = \theta^*$

$$\nabla\Phi(\theta^*) = \left(-\frac{\partial^2 Q}{\partial \theta'^2}(\theta^*|\theta^*)\right)^{-1} \frac{\partial^2 Q}{\partial \theta \partial \theta'}(\theta^*|\theta^*) \tag{1}$$

- If $\frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t+1)}|\theta^{(t)})$ is negative definite with eigenvalues bounded away from zero, then

$$-\frac{\partial^2 Q}{\partial \theta'^2}(\theta^*|\theta^*) = \mathbb{E}_{p(z|x,\theta^*)}\left(-\nabla^2 \log p(x,z|\theta^*)\right)$$

is positive definite, known as the **complete information**

- The marginal log-likelihood can be rewritten as

$$\begin{aligned}\mathcal{L}(\theta') &= \mathbb{E}_{p(z|x,\theta)} \log p(x,z|\theta') - \mathbb{E}_{p(z|x,\theta)} \log p(z|x,\theta) \\ &= Q(\theta'|\theta) - H(\theta'|\theta)\end{aligned}$$

Therefore

$$\frac{\partial^2 Q}{\partial \theta \partial \theta'}(\theta'|\theta) = \frac{\partial^2 H}{\partial \theta \partial \theta'}(\theta'|\theta)$$

- Some properties of $H(\theta|\theta) = \mathbb{E}_{p(z|x,\theta)} \log p(z|x,\theta)$

$$\frac{\partial H}{\partial \theta'}(\theta|\theta) = 0$$
$$\frac{\partial^2 H}{\partial \theta \partial \theta'}(\theta|\theta) = -\frac{\partial^2 H}{\partial \theta'^2}(\theta|\theta)$$

- Therefore,

$$\frac{\partial^2 Q}{\partial \theta \partial \theta'}(\theta^*|\theta^*) = \frac{\partial^2 H}{\partial \theta \partial \theta'}(\theta^*|\theta^*) = -\frac{\partial^2 H}{\partial \theta'^2}(\theta^*|\theta^*)$$

is positive semidefinite (variance of the score $\nabla \log p(z|x,\theta^*)$), known as the **missing information**

$$\mathcal{L}(\theta') = Q(\theta'|\theta) - H(\theta'|\theta)$$

- Differentiate both side w.r.t. θ' twice

$$\nabla^2\mathcal{L}(\theta') = \frac{\partial^2 Q}{\partial\theta'^2}(\theta'|\theta) - \frac{\partial^2 H}{\partial\theta'^2}(\theta'|\theta)$$

- The *missing-information principle*

$$\underbrace{-\frac{\partial^2 Q}{\partial\theta'^2}(\theta|\theta)}_{I_{\text{complete}}} = \underbrace{-\nabla^2\mathcal{L}(\theta)}_{I_{\text{observed}}} + \underbrace{-\frac{\partial^2 H}{\partial\theta'^2}(\theta|\theta)}_{I_{\text{missing}}}$$

- Substitute in (1)

$$\begin{aligned}\nabla\Phi(\theta^*) &= I_{\text{complete}}^{-1}(\theta^*) I_{\text{missing}}(\theta^*) \\ &= \left(I_{\text{observed}}(\theta^*) + I_{\text{missing}}(\theta^*)\right)^{-1} I_{\text{missing}}(\theta^*)\end{aligned}$$

- When $I_{\text{observed}} = -\nabla^2 \mathcal{L}(\theta^*)$ is positive definite, the eigenvalues of $\nabla\Phi(\theta^*)$ are all less than 1, EM has a linear convergence rate.
- The rate of convergence depends on the relative size of $I_{\text{observed}}(\theta^*)$ and $I_{\text{missing}}(\theta^*)$. EM converges rapidly when the missing information is small.
- The fraction of information loss may vary across different component of θ, so some component may converge faster than other components.
- See Wu (1983) for more detailed discussions.

EM for Maximum A Posterior

- EM can be easily modified for the Maximum A Posterior (MAP) estimate instead of the MLE.
- Suppose the log-prior penalty term is $R(\theta)$. We only have to maximize

$$Q(\theta|\theta^{(t)}) + R(\theta) \tag{2}$$

  in the M-step
- Monotonicity.

$$\begin{aligned}
\mathcal{L}(\theta^{(t+1)}) + R(\theta^{(t+1)}) &\geq \mathcal{F}(\theta^{(t+1)}|\theta^{(t)}) + R(\theta^{(t+1)}) \\
&\geq \mathcal{F}(\theta^{(t)}|\theta^{(t)}) + R(\theta^{(t)}) \\
&= \mathcal{L}(\theta^{(t)}) + R(\theta^{(t)})
\end{aligned}$$

- If $R(\theta)$ corresponds to conjugate prior, (2) can be maximized in the same manner as $Q(\theta|\theta^{(t)})$.

- The E-step requires finding the expected complete data log-likelihood $Q(\theta|\theta^{(t)})$. When this expectation is difficult to compute, we can approximate it via Monte Carlo methods
- **Monte Carlo EM** (Wei and Tanner, 1990)
  - Draw missing data $z_1^{(t)}, \ldots, z_m^{(t)}$ from the conditional distribution $p(z|x, \theta^{(t)})$
  - Compute a Monte Carlo estimate of $Q(\theta|\theta^{(t)})$

$$\hat{Q}^{(t+1)}(\theta|\theta^{(t)}) = \frac{1}{m}\sum_{i=1}^{m} \log p(x, z_i^{(t)}|\theta)$$

  - Update $\theta^{(t+1)}$ to maximize $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})$.

Remark: It is recommended to let m changes along iterations (small at the beginning and increases as iterations progress)

- By the lack of memory, it is easy to compute the expected complete data log-likelihood, which lead to the ordinary EM update

$$\mu_{\mathrm{EM}}^{(k+1)} = \frac{\sum_{i=1}^{n} t_i + (n-r)\mu^{(k)}}{n}$$

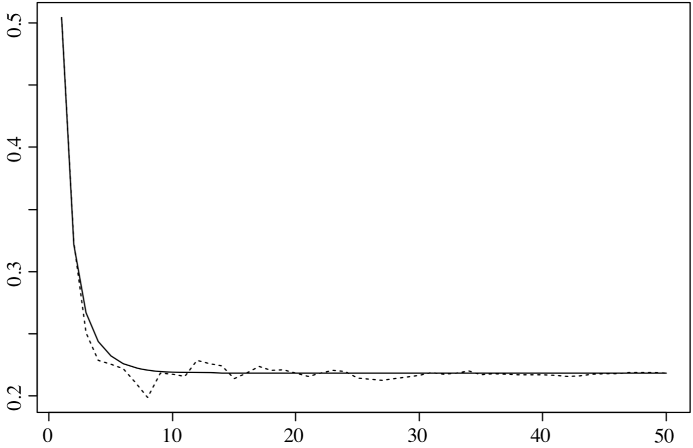
- In MCEM, we can sample from the conditional distribution

$$\boldsymbol{T}_j = (T_{j,r+1}, \ldots, T_{j,n}),\ T_{j,l} - t_l \sim \mathrm{Exp}(\mu^{(k)}), \quad l = r+1, \ldots, n$$

for $j = 1, \ldots, m^{(k)}$, and the update formula is

$$\mu_{\mathrm{MCEM}}^{(k+1)} = \frac{\sum_{i=1}^{n} t_i + \frac{1}{m^{(k)}} \sum_{j=1}^{m^{(k)}} \boldsymbol{T}_j^T \mathbf{1}}{n}$$

Examples: Censored Survival Times



Improving the M-step

- One of the appeals of the EM algorithm is that $Q(\theta|\theta^{(t)})$ is often simpler to maximize than the marginal likelihood
- In some cases, however, the M-step cannot be carried out easily even though the computation of $Q(\theta|\theta^{(t)})$ is straightforward in the E-step
- For such situations, Dempster et al (1977) defined a generalized EM algorithm (GEM) for which the M-step only requires $\theta^{(t+1)}$ to improve $Q(\theta|\theta^{(t)})$

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t+1)}|\theta^{(t)})$$

- We can easily show that GEM is also monotonic in $\mathcal{L}$

$$\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{F}(q^{(t)}, \theta^{(t+1)}) \geq \mathcal{F}(q^{(t)}, \theta^{(t)}) = \mathcal{L}(\theta^{(t)})$$

Expectation Conditional Maximization

- Meng and Rubin (1993) replaces the M-step with a series of computationally cheaper conditional maximization (CM) steps, leading to the **ECM** algorithm
- The M-step in ECM contains a collection of simple CM steps, called a CM *cycle*. For $s = 1, \ldots, S$, the s-th CM step requires the maximization of $Q(\theta|\theta^{(t)})$ subject to a constraint

$$\theta^{(t+s/S)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}), \quad \text{s.t. } g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$$

- The efficiency of ECM depends on the choice of constraints. Examples: Blockwise updates (coordinate ascent).
- One may also insert an E-step between each pair of CM-steps, updating Q at every stage of the CM cycle.

- Suppose we have n independent observations from the following k-variate normal model

$$Y_i \sim \mathcal{N}(X_i\beta, \Sigma), \quad i = 1, \ldots, n$$

  - $X_i \in \mathbb{R}^{k\times p}$ is a known design matrix for the i-th observation
  - β is a vector of p unknown parameters
  - Σ is a $d \times d$ unknown variance-covariance matrix
- The complete data log-likelihood (up to a constant) is

$$L(\beta, \Sigma|Y) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta)$$

- Generally, MLE does not has closed form solution except in special cases (e.g., $\Sigma = \sigma^2 I$)

A Coordinate Ascent Algorithm

- Although the joint maximization of β and Σ are not generally in closed form, a coordinate ascent algorithm does exist
- Given $\Sigma = \Sigma^{(t)}$, the conditional MLE of β is simply the weighted least-square estimate

$$\beta^{(t+1)} = \left(\sum_{i=1}^{n} X_i^T (\Sigma^{(t)})^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^T (\Sigma^{(t)})^{-1} Y_i\right)$$

- Given $\beta = \beta^{(t+1)}$, the conditional MLE of Σ is the cross-product of the residuals

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\beta^{(t+1)})(Y_i - X_i\beta^{(t+1)})^T$$

- Now suppose that we also have missing data

$$Y_i \sim \mathcal{N}(X_i\beta, \Sigma), \quad i = n+1, \ldots, m$$

  for which only the design matrix $X_i,\ i > n$ are known
- The complete data log-likelihood

$$L(\beta, \Sigma|Y) = -\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta)$$

- Expected values of sufficient statistics observed data and current parameter $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)})$

$$\mathbb{E}(Y_i|Y_{\text{obs}}, \theta^{(t)}) = X_i\beta^{(t)}$$
$$\mathbb{E}(Y_iY_i^T|Y_{\text{obs}}, \theta^{(t)}) = \Sigma^{(t)} + (X_i\beta^{(t)})(X_i\beta^{(t)})^T$$

E-step

Expected complete-data log-likelihood

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) = &-\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta) \\
&- \frac{1}{2}\sum_{i=n+1}^{m}\mathbb{E}\left((Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta)\right) \\
= &-\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta) \\
&- \frac{1}{2}\sum_{i=n+1}^{m}(\mathbb{E}Y_i - X_i\beta)^T\Sigma^{-1}(\mathbb{E}Y_i - X_i\beta) + C
\end{aligned}
$$

where $C = \frac{1}{2}\sum_{i=n+1}^{m}\mathbb{E}(Y_i)^T\Sigma^{-1}\mathbb{E}(Y_i) - \mathbb{E}(Y_i^T\Sigma^{-1}Y_i)$ is a constant independent of the parameter β.

CM-steps: Update β

- The first CM-step, maximize Q given $\Sigma = \Sigma^{(t)}$.
- Since C is independent of β, we can maximize

$$
\begin{aligned}
-\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta) \\
-\frac{1}{2}\sum_{i=n+1}^{m}(\mathbb{E}Y_i - X_i\beta)^T\Sigma^{-1}(\mathbb{E}Y_i - X_i\beta)
\end{aligned}
$$

$$
\Rightarrow \beta^{(t+1)} = \left(\sum_{i=1}^{m} X_i^T\Sigma^{(t)}X_i\right)^{-1}\left(\sum_{i=1}^{m} X_i^T\Sigma^{(t)}\hat{Y}_i\right)
$$

where

$$
\hat{Y}_i = \begin{cases} Y_i, & i \le n \\ X_i\beta^{(t)}, & i > n \end{cases}
$$

CM-steps: Update Σ

- The second CM-step, maximize Q with $\beta = \beta^{(t+1)}$
- Rewrite Q as

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) = \frac{m}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left(\Sigma^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^T\right) \\
- \frac{1}{2}\sum_{i=n+1}^{m}\mathrm{Tr}\left(\Sigma^{-1}\mathbb{E}\left((Y_i - X_i\beta)(Y_i - X_i\beta)^T\right)\right)
\end{aligned}
$$

- Similarly as in the complete data case

$$
\begin{aligned}
\Sigma^{(t+1)} =& \frac{1}{m}\left(\sum_{i=1}^{n}(Y_i - X_i\beta^{(t+1)})(Y_i - X_i\beta^{(t+1)})^T + \sum_{i=n+1}^{m}\Sigma^{(t)}\right. \\
&\left. + \sum_{i=n+1}^{m}X_i(\beta^{(t)} - \beta^{(t+1)})(\beta^{(t)} - \beta^{(t+1)})^T X_i^T\right)
\end{aligned}
$$

- Both the E-step and the two CM-steps can be implemented using close form solutions, no numerical iteration required.
- Both CM-steps improves Q

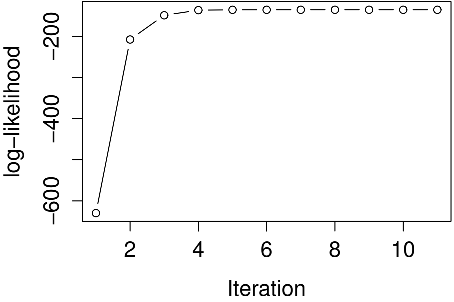
$$
\begin{aligned}
Q(\beta^{(t+1)}, \Sigma^{(t+1)} | \beta^{(t)}, \Sigma^{(t)}) &\geq Q(\beta^{(t+1)}, \Sigma^{(t)} | \beta^{(t)}, \Sigma^{(t)}) \\
&\geq Q(\beta^{(t)}, \Sigma^{(t)} | \beta^{(t)}, \Sigma^{(t)})
\end{aligned}
$$

- ECM in this case can be viewed as an efficient generalization of iterative reweighted least squares, in the presence of missing data.

We generate 120 design matrices at random and simulate 100 observations with $\beta = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1, & 0.1 \\ 0.1 & 2 \end{pmatrix}$

ECM estimates

$$\hat{\beta} = \begin{pmatrix} 2.068 \\ 1.087 \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} 0.951 & 0.214 \\ 0.214 & 2.186 \end{pmatrix}$$



EM Gradient Algorithm

- Iterative optimization can be considered when direct maximization is not available.
- All numerical optimization can apply and that would yield an algorithm that has nested iterative loops (e.g., ECM inserts conditional maximization steps within each CM cycle)
- To avoid the computational burden of nested looping, Lange proposed to use one single step of Newton's method

$$
\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} - \left(\frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t)}|\theta^{(t)})\right)^{-1} \frac{\partial Q}{\partial \theta'}(\theta^{(t)}|\theta^{(t)}) \\
&= \theta^{(t)} - \left(\frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t)}|\theta^{(t)})\right)^{-1} \nabla \mathcal{L}(\theta^{(t)})
\end{aligned}
$$

- This EM gradient algorithm has the same rate of convergence as the full EM algorithm.

- When EM is slow, we can use the relatively simple analytic setup from EM to motivate particular forms for Newton-like steps.
- **Aitken Acceleration**. Newton update

$$\theta^{(t+1)} = \theta^{(t)} - (\nabla^2 \mathcal{L}(\theta^{(t)}))^{-1} \nabla \mathcal{L}(\theta^{(t)}) \tag{3}$$

  Note that $\nabla \mathcal{L}(\theta^{(t)}) = \frac{\partial Q}{\partial \theta'}(\theta^{(t)}|\theta^{(t)})$ and

$$0 = \frac{\partial Q}{\partial \theta'}(\theta_{\mathrm{EM}}^{(t+1)}|\theta^{(t)}) \approx \frac{\partial Q}{\partial \theta'}(\theta^{(t)}|\theta^{(t)}) + \frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t)}|\theta^{(t)})(\theta_{\mathrm{EM}}^{(t+1)} - \theta^{(t)})$$

  substitute in (3)

$$\theta^{(t+1)} = \theta^{(t)} + (I_{\mathrm{observed}}(\theta^{(t)}))^{-1} I_{\mathrm{complete}}(\theta^{(t)})(\theta_{\mathrm{EM}}^{(t+1)} - \theta^{(t)})$$

- Many other acceleration exists (e.g., Quasi-Newton methods).

References

- A. P. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39:1–38, 1977.
- R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models, 89:355–368, 1998.
- Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In Shisha, O. (Ed.), Inequalities III: Proceedings of the 3rd Symposium on Inequalities, 1–8. Academic Press.

References

- C. F. J. Wu. On the convergence properties of the EM algorithm. Annals of Statistics, 11:95–103, 1983.
- X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika, 80:267–278, 1993.
- G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association, 85:699–704, 1990.
- K. Lange. A gradient algorithm locally equivalent to the EM algorithm. Journal of the Royal Statistical Society, Series B, 57:425–437, 1995.

References

- T. A. Louis. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B, 44:226–233, 1982.