STA2311: Advanced Computational Methods for Statistics I

Class 3: The EM Algorithm

Radu Craiu   Robert Zimmerman

University of Toronto

September 26, 2023

Section 1

Introduction

Background

- The *expectation-maximization (EM) algorithm* is one of the most ubiquitous algorithms in statistics
- It plays a huge role in both frequentist and Bayesian computational statistics
- The algorithm was formally introduced in Dempster et al. [1977] but a number of special cases of it were known earlier
  - e.g., the *Baum-Welch algorithm* for fitting hidden Markov models
- In the original setup, the goal is to find the MLE of some parameter $\boldsymbol{\theta}$ in a statistical model featuring missing data

Section 2

Missing Data and the Algorithm Itself

Missing Data?

- It is not unusual for data to be lost or unreported
- The EM algorithm is thus popular in many areas
- In addition, it can be helpful to formulate a model with complete data as a missing data one (as we will see later)
- In fact, some of the most commonly-used statistical models benefit from the elegant properties of the EM algorithm

The Basics

- Suppose that under perfect circumstances, we could observe *complete data* $\tilde{\boldsymbol{Y}}_{\text{com}}$ generated from some statistical model
- In real life, however, we only have access to a part of it, the *observed data* $\tilde{\boldsymbol{Y}}_{\text{obs}}$
- The remaining part is the *missing data* $\tilde{\boldsymbol{Y}}_{\text{mis}}$
- So $\tilde{\boldsymbol{Y}}_{\text{com}} = (\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}})$

Missingness Mechanisms

- Let $f(\tilde{\boldsymbol{y}}_{\text{com}} \mid \boldsymbol{\theta})$ and $g(\tilde{\boldsymbol{y}}_{\text{obs}} \mid \boldsymbol{\theta})$ be the pdfs of the complete and observed data, respectively
- Define the random variable R as

$$R = \begin{cases} 1 & \text{if } \boldsymbol{Y}_{\text{mis}} \text{ is observed} \\ 0 & \text{if } \boldsymbol{Y}_{\text{mis}} \text{ is unobserved} \end{cases}$$

- Suppose the distribution of R depends on $\boldsymbol{Y}_{\text{com}}$ and varies with some parameter ψ
  - i.e., it takes the form $p(r \mid \boldsymbol{Y}_{\text{com}}, \psi)$
- The likelihood of the model that includes the missing indicator is then

$$L(\boldsymbol{\theta}, \psi \mid \tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{R}) = \int p(\tilde{R} \mid \tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}}, \psi) f(\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \boldsymbol{\theta}) \, \mathrm{d}\tilde{\boldsymbol{Y}}_{\text{mis}}$$

MAR and MCAR

- Suppose that the probability of missingness does not depend on the missing data itself
  - i.e., $p(R \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}, \psi) = p(R \mid \boldsymbol{Y}_{\text{obs}}, \psi)$
  - This property is known as *missing at random (MAR)*
- Then

$$L(\boldsymbol{\theta}, \psi \mid \tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{R}) = p(\tilde{R} \mid \tilde{\boldsymbol{Y}}_{\text{obs}}, \psi) g(\tilde{\boldsymbol{Y}}_{\text{obs}} \mid \theta)$$

- So in this case, likelihood-based inference for $\boldsymbol{\theta}$ does not depend on the distribution of R
- So we can proceed without considering the missingness mechanism
- The stronger condition $p(R \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}, \psi) = p(R \mid \psi)$ is known as *missing completely at random (MCAR)*

The Ingredients

- Define the *complete-data log likelihood* as

$$\ell_{\text{com}}(\boldsymbol{\theta}) = \log\Big(f(\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \boldsymbol{\theta})\Big)$$

  and the *observed-data log likelihood* as

$$\ell_{\text{obs}}(\boldsymbol{\theta}) = \log\Big(g(\tilde{\boldsymbol{Y}}_{\text{obs}} \mid \boldsymbol{\theta})\Big)$$

- Define the *Q-function* as the conditional expectation

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{\theta}'}\Big[\ell_{\text{com}}(\boldsymbol{\theta}) \mid \tilde{\boldsymbol{Y}}_{\text{obs}}\Big]$$

  computed with respect to the conditional density

$$k(\tilde{\boldsymbol{Y}}_{\text{mis}} \mid \tilde{\boldsymbol{Y}}_{\text{obs}}, \boldsymbol{\theta}') = \frac{f(\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \boldsymbol{\theta}')}{g(\tilde{\boldsymbol{Y}}_{\text{obs}} \mid \boldsymbol{\theta}')}$$

Description of the EM Algorithm

- The EM algorithm relies on an iterative procedure to find a local (or global) maximizer of $\ell_{\text{obs}}(\boldsymbol{\theta})$:
  1. Choose a starting value $\boldsymbol{\theta}^{(0)}$
  2. For $t \geq 0$:
     - E-Step Compute $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$
     - M-Step Set $\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$
  3. Stop when $\frac{||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}||}{||\boldsymbol{\theta}^{(t)}||} < \epsilon$ where ϵ is a small user-defined threshold (say $\epsilon \approx 10^{-6}$)

Section 3

Examples

Example: Finite Mixture of Poissons

- Suppose $Y_1, \ldots, Y_n$ arises from a finite mixture of K Poisson distributions:

$$\mathbb{P}(Y_i = y) = \sum_{k=1}^{K} \pi_k \frac{\lambda_k^{y_i} e^{-\lambda_k}}{y_i!}$$

- Here each $\lambda_k > 0$, and each $\pi_k > 0$ with $\sum_{k=1}^{K} \pi_k = 1$
- We want to find the MLE of $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_K, \pi_1, \ldots, \pi_K)$
  - Note that $\pi_K = 1 - \pi_1 - \cdots - \pi_{K-1}$, so there are $2K - 1$ scalars to estimate
- The MLE does not exist in closed form, but the model is an ideal candidate for the EM algorithm

Example: Finite Mixture of Poissons (Continued)

- To begin with, we need to formulate the latent variables and the complete data
- It is easy to show that the original model is equivalent to the model

$$
\begin{aligned}
Y_i \mid Z_i = k &\sim \text{Poisson}(\lambda_k) \\
Z_i &\sim \text{Categorical}(\pi_1, \ldots, \pi_k)
\end{aligned}
$$

- That is, $\mathbb{P}(Y_i = y \mid Z_i = k) = \frac{\lambda_k^{y_i} e^{-\lambda_k}}{y_i!}$ and $\mathbb{P}(Z_i = k) = \pi_k$ for $1 \leq k \leq K$
- So we take $\tilde{\boldsymbol{Y}}_{\text{com}} = (Y_1, \ldots, Y_n, Z_1, \ldots, Z_n)$ and $\tilde{\boldsymbol{Y}}_{\text{mis}} = (Z_1, \ldots, Z_n)$

Example: Finite Mixture of Poissons (Continued)

- The pdf of the complete data is given by

$$f(\tilde{\boldsymbol{Y}}_{\text{com}} \mid \boldsymbol{\theta}) = f(\tilde{\boldsymbol{y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\pi_k \cdot \frac{\lambda_k^{y_i} e^{-\lambda_k}}{y_i!} \right)^{\mathbb{1}_{Z_i = k}}$$

- The complete-data log likelihood is therefore

$$\ell_{\text{com}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{Z_i = k} \cdot (\log(\pi_k) + y_i \cdot \log(\lambda_k) - \lambda_k - \log(y_i!))$$

- Taking the expectation with respect to $\boldsymbol{\theta}'$ and conditional on $\tilde{\boldsymbol{Y}}_{\text{obs}} = \tilde{\boldsymbol{y}}_{\text{obs}}$, our Q-function is

$$\begin{aligned} & Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') \\ = & \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}(Z_i = k \mid \tilde{\boldsymbol{y}}_{\text{obs}}, \boldsymbol{\theta}') \cdot (\log(\pi_k) + y_i \cdot \log(\lambda_k) - \lambda_k - \log(y_i!)) \end{aligned}$$

Example: Finite Mixture of Poissons (Continued)

- To evaluate $\mathbb{P}(Z_i = k \mid \tilde{\boldsymbol{Y}}_{\text{obs}} = \tilde{\boldsymbol{y}}_{\text{obs}}, \boldsymbol{\theta}')$, use Bayes' rule and the law of total probability to get

$$\begin{aligned}
\mathbb{P}(Z_i = k \mid \tilde{\boldsymbol{Y}}_{\text{obs}} = \tilde{\boldsymbol{y}}_{\text{obs}}, \boldsymbol{\theta}') &= \frac{\mathbb{P}(\tilde{\boldsymbol{Y}}_{\text{obs}} = \tilde{\boldsymbol{y}}_{\text{obs}} \mid Z_i = k, \boldsymbol{\theta}') \cdot \mathbb{P}(Z_i = k \mid \boldsymbol{\theta}')}{\sum_{l=1}^{K} \mathbb{P}(\tilde{\boldsymbol{Y}}_{\text{obs}} = \tilde{\boldsymbol{y}}_{\text{obs}} \mid Z_i = l, \boldsymbol{\theta}') \cdot \mathbb{P}(Z_i = l \mid \boldsymbol{\theta}')} \\
&= \frac{\mathbb{P}(Y_i = y_i \mid Z_i = k, \boldsymbol{\theta}') \cdot \mathbb{P}(Z_i = k \mid \boldsymbol{\theta}')}{\sum_{l=1}^{K} \mathbb{P}(Y_i = y_i \mid Z_i = l, \boldsymbol{\theta}') \cdot \mathbb{P}(Z_i = l \mid \boldsymbol{\theta}')} \\
&= \frac{\pi'_k \cdot \frac{\lambda_k'^{y_i} e^{-\lambda'_k}}{y_i!}}{\sum_{l=1}^{K} \pi'_l \cdot \frac{\lambda_l'^{y_i} e^{-\lambda'_l}}{y_i!}} \\
&=: a_k(y_i, \boldsymbol{\theta}')
\end{aligned}$$

- So

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = \sum_{i=1}^{n} \sum_{k=1}^{K} a_k(y_i, \boldsymbol{\theta}') \cdot (\log(\pi_k) + y_i \cdot \log(\lambda_k) - \lambda_k - \log(y_i!))$$

Example: Finite Mixture of Poissons (Continued)

- We must now maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ in $\boldsymbol{\theta}$, which amounts to finding $\nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$, setting it to $\mathbf{0}$, and solving
- Basic calculus and some algebra shows that the maximizing parameters are given by

$$\hat{\lambda}_k = \frac{\sum_{i=1}^n y_i \cdot a_k(y_i, \boldsymbol{\theta}')}{\sum_{i=1}^n a_k(y_i, \boldsymbol{\theta}')}, \quad 1 \leq k \leq K$$

and

$$\hat{\pi}_k = \frac{\sum_{i=1}^n a_k(y_i, \boldsymbol{\theta}')}{\sum_{l=1}^K \sum_{i=1}^n a_l(y_i, \boldsymbol{\theta}')}, \quad 1 \leq k \leq K$$

Example: Finite Mixture of Poissons (Continued)

- The EM algorithm for this example is thus
  1. Choose a starting value $\boldsymbol{\theta}^{(0)}$
  2. For $t \geq 0$: compute $\boldsymbol{\theta}^{(t)}$ via the updates

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n y_i \cdot a_k(y_i, \boldsymbol{\theta}^{(t)})}{\sum_{i=1}^n a_k(y_i, \boldsymbol{\theta}^{(t)})}$$

     and

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^n a_k(y_i, \boldsymbol{\theta}^{(t)})}{\sum_{l=1}^K \sum_{i=1}^n a_l(y_i, \boldsymbol{\theta}^{(t)})}$$

  3. Stop when $\frac{||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}||}{||\boldsymbol{\theta}^{(t)}||} < \epsilon$ where ϵ is a small user-defined threshold (say $\epsilon \approx 10^{-6}$)

Example: Finite Mixture of Poissons (Continued)

```
set.seed(2311)
norm <- function(x) {sqrt(sum(x^2))}

n <- 10000
lambda_true <- c(0.5, 2.5, 5)
y <- rep(0, times=n)

for (i in 1:n) {
  z <- which(rmultinom(n=1, size=c(1,1,1), prob=c(0.25, 0.5, 0.25)) == 1)
  y[i] <- rpois(n=1, lambda=lambda_true[z])
}
```

Example: Finite Mixture of Poissons (Continued)

```
lambda_new = c(0.5, 1, 1.5)
pi_new <- c(1/3, 1/3, 1/3)
theta_new <- c(lambda_new, pi_new)
theta_old <- rep(1000, times=3)

A <- array(0L, dim=c(3, n))

while(norm(theta_new - theta_old)/norm(theta_old) >= 1e-6) {
  theta_old <- theta_new; pi_old <- pi_new; lambda_old <- lambda_new
  for (k in 1:3) {
    for (i in 1:n) {
      A[k, i] <- pi_old[k]*dpois(y[i], lambda_old[k])/
        sum(pi_old*dpois(y[i], lambda_old))
    }
  }

  lambda_new <- sapply(1:3, function(k) sum(y*A[k,])/sum(A[k,]))
  pi_new <- sapply(1:3, function(k) sum(A[k,])/sum(A))
  theta_new <- c(lambda_new, pi_new)
  print(theta_new)
}
```

Exponential Families

- Recall that the distribution of a random vector $\boldsymbol{Y}$ is in an *exponential family* if its density (or mass) function can be written as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}) = h(\boldsymbol{y}) \cdot g(\boldsymbol{\theta}) \cdot \exp\Big(\eta(\boldsymbol{\theta})^\top T(\boldsymbol{y})\Big),$$

  where $T(\cdot)$, $\eta(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are known functions
- $T(\boldsymbol{Y})$ is called the *sufficient statistic* for the distribution
- Exponential families have countless properties that make them particularly nice to do inference with
- Many "classical" distributions are members of exponential families
  - Normal, exponential, chi-squared, gamma, beta, Bernoulli, binomial, negative binomial, multinomial, Poisson, geometric...
  - Finite mixtures of exponential family distributions don't count

EM for Exponential Families

- If the distribution of $\boldsymbol{Y}_{\text{com}}$ is in an exponential family, then the EM algorithm has a particularly simple form
- At the t'th iteration:

E-Step Estimate the sufficient statistic $T = T(\boldsymbol{Y}_{\text{com}})$ by

$$T^{(t)} = \mathbb{E}[T(\boldsymbol{Y}_{\text{com}}) \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}]$$

M-Step Compute $\boldsymbol{\theta}^{(t+1)}$ by solving

$$\mathbb{E}\left[\frac{\partial \eta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{\top} T(\boldsymbol{Y}_{\text{com}}) \mid \boldsymbol{\theta}\right] = \frac{\partial \eta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{\top} T^{(t)},$$

or, if the Jacobian $\frac{\partial \eta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is invertible, by solving

$$\mathbb{E}[T(\boldsymbol{Y}_{\text{com}}) \mid \boldsymbol{\theta}] = T^{(t)}$$

EM for Bayesian Posteriors

- Suppose we're Bayesians and we equip $\boldsymbol{\theta}$ with a prior $p(\boldsymbol{\theta})$
- Instead of the MLE, we want to find the posterior mode $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}) \cdot g(\tilde{\boldsymbol{y}}_{\text{obs}} \mid \boldsymbol{\theta})$
- Fortunately, the EM algorithm can handle this
- Instead of maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ in the M-step, we simply maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) + p(\boldsymbol{\theta})$
- All of the theory still works!

Section 4

Convergence Properties

EM: The Ascent Property

- A primary feature of the EM algorithm is that each new iterate $\boldsymbol{\theta}^{(t+1)}$ never decreases the likelihood from the previous one:

Theorem

Let $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(t)}, \ldots$ be the sequence of parameter estimates produced by the EM algorithm. For all $t \geq 0$,

$$L(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{Y}_{obs}) \geq L(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{Y}_{obs}).$$

- This is not hard to prove
- First note that

$$\begin{aligned}\ell(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{Y}_{\text{obs}}) &= \log\Big(g(\boldsymbol{Y}_{\text{obs}} \mid \boldsymbol{\theta}^{(t+1)})\Big) \\ &= \ell(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{Y}_{\text{com}}) - \log\Big(k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)})\Big)\end{aligned}$$

EM: The Ascent Property (Continued)

- Taking expectations with respect to $\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}$, we obtain

$$
\begin{aligned}
&\ell(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{Y}_{\text{obs}}) \\
&= Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)})\right) \mid \boldsymbol{Y}_{\text{obs}}\right] \\
&\geq Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)})\right) \mid \boldsymbol{Y}_{\text{obs}}\right] \\
&\geq Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)})\right) \mid \boldsymbol{Y}_{\text{obs}}\right] \\
&= \ell(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{Y}_{\text{obs}})
\end{aligned}
$$

- The first inequality is true because $\boldsymbol{\theta}^{(t+1)}$ is chosen to maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$

EM: The Ascent Property (Continued)

- The second inequality is essentially due to Jensen's inequality:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)})\right) \mid \boldsymbol{Y}_{\text{obs}}\right] \\
&\qquad\qquad - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)})\right) \mid \boldsymbol{Y}_{\text{obs}}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(\frac{k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)})}{k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)})}\right) \mid \boldsymbol{Y}_{\text{obs}}\right] \\
&\leq \log\left(\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\frac{k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)})}{k(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)})} \mid \boldsymbol{Y}_{\text{obs}}\right]\right) \\
&= \log\left(\int \frac{k(\boldsymbol{y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)})}{k(\boldsymbol{y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)})} k(\boldsymbol{y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}) \, \mathrm{d}\boldsymbol{y}_{\text{mis}}\right) \\
&= \log\left(\int k(\boldsymbol{y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t+1)}) \, \mathrm{d}\boldsymbol{y}_{\text{mis}}\right) \\
&= 0.
\end{aligned}
$$

Generalized EM Algorithms

- The proof above shows that the theorem holds for any sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(t)}, \ldots$ such that $Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})$ for all $t \geq 0$
- Algorithms which produce such sequences are known as *generalized EM algorithms*
- These are also described in Dempster et al. [1977]
- A famous example is the *ECM algorithm* of Meng and Rubin [1993]
  - This essentially updates $\boldsymbol{\theta}$ one (or several) components at a time within the M-step
  - A further extension is the *ECME algorithm* of Liu and Rubin [1994], which speeds up the ECM algorithm

Initialization(s)

- The ascent property shows that the generalized EM algorithms will eventually find a local maximum of the log-likelihood function (if one exists)
- But there is no guarantee that this is the global maximum!
- Likelihood functions for complicated models with many parameters may have many local maxima, and the algorithm may become stuck in one
- Thus, it is usually a good idea to run the algorithm several times with different initial values
- If the parameter estimates upon convergence appear robust to initial values, we have more assurance that the algorithm has discovered the global maximum

Section 5

Variance Calculations and Convergence Rates

Asymptotic Variance of the MLE

- Classical theory tells us that under certain regularity conditions, the MLE $\boldsymbol{\theta}_{\text{MLE}}$ for a statistical model $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is asymptotically normal
- The asymptotic covariance is usually estimated using the inverse of the observed information, $\boldsymbol{\mathcal{I}}_{\text{obs}}(\boldsymbol{\theta}_{\text{MLE}}) := \left[-\boldsymbol{H}_{\ell}(\boldsymbol{\theta})\,\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MLE}}}\right]^{-1}$
  - Here $\boldsymbol{H}_{\ell}(\boldsymbol{\theta})$ is the negative Hessian of the log-likelihood, as a function of $\boldsymbol{\theta}$
- However, the Hessian is generally unavailable when using the EM algorithm to find $\boldsymbol{\theta}_{\text{MLE}}$
- Usually, the complete data version of the observed information is easier to compute than that based on the marginal likelihood

Louis's Method

- Recall from the proof of the ascent property that

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - R(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}),$$

  where

$$R(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log(k(\boldsymbol{Y}_{\text{mis}}) \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}) \mid \boldsymbol{Y}_{\text{obs}}]$$

- Suppose the EM algorithm has terminated, so that $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ is the MLE (or a stationary point of the algorithm)
- Taking negative second derivatives of both sides gives

$$\mathcal{I}_{\text{obs}}(\boldsymbol{\theta}) = \mathcal{I}_{\text{com}}(\boldsymbol{\theta}) - \mathcal{I}_{\text{mis}}(\boldsymbol{\theta}), \tag{1}$$

  where $\mathcal{I}_{\text{com}}(\boldsymbol{\theta}) = -\boldsymbol{H}_{Q(\cdot|\boldsymbol{\theta}^*)}(\boldsymbol{\theta})$ is called the *complete information* and $\mathcal{I}_{\text{mis}}(\boldsymbol{\theta}) = \boldsymbol{H}_{R(\cdot|\boldsymbol{\theta}^*)}(\boldsymbol{\theta})$ is called the *missing information*

The Missing Information Principle

- What happens when we evaluate (1) at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$?
- To simplify notation, assume that θ is a scalar
  - Everything extends to vector parameters when first derivatives are replaced by gradients and second derivatives are replaced by Hessians
- Under regularity conditions, the complete information evaluated at θ^* can be written as

$$\mathcal{I}_{\text{com}}(\theta^*) = \mathbb{E}_{\theta^*}\left[-\frac{\partial^2}{\partial \theta^2} \log\Big(f(\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \theta)\Big)\Big|_{\theta=\theta^*} \mid \tilde{\boldsymbol{Y}}_{\text{obs}} \right]$$

- Similarly, the missing information at θ^* can be written as

$$\begin{aligned}\mathcal{I}_{\text{mis}}(\theta^*) &= \mathbb{E}_{\theta^*}\left[\left(\frac{\partial}{\partial \theta} \log\Big(f(\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \theta)\Big)\right)^2 \Big|_{\theta=\theta^*} \mid \tilde{\boldsymbol{Y}}_{\text{obs}} \right] \\ &= \mathrm{Var}_{\theta^*}\left(\frac{\partial}{\partial \theta} \log\Big(f(\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \theta)\Big)\Big|_{\theta=\theta^*} \mid \tilde{\boldsymbol{Y}}_{\text{obs}} \right)\end{aligned}$$

The Missing Information Principle (Continued)

- So the missing information is the conditional variance of the complete-data score function, and is always non-negative
- More missing data will result in a larger reduction of the observed information
- Hence, the asymptotic variance (i.e., $\mathcal{I}_{\text{com}}^{-1}(\boldsymbol{\theta}^*)$) will be larger
- This is not surprising, as we expect to obtain estimators with larger variances when data are missing
- The same principle is also intimately connected to the algorithm's rate of convergence

Rate of Convergence

- An optimization method for finding $\boldsymbol{\theta}^*$ with convergence order c has a *rate of convergence* γ if $\lim_{t\to\infty} \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ and

$$\lim_{t\to\infty} \frac{||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*||}{||\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*||^c} = \gamma,$$

  provided the limit exists

- The convergence order of the EM algorithm is usually 1 (i.e., it converges linearly)
  - In contrast to, e.g., Newton-Raphson, which is quadratic but lacks the ascent property

- If the EM update is implicitly defined by the function $\boldsymbol{M}(\cdot)$ (i.e., $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{M}(\boldsymbol{\theta}^{(t)})$), then the EM algorithm's rate of convergence is given by the largest eigenvalue of the Jacobian $\frac{\partial M}{\partial \boldsymbol{\theta}}$

The Fraction of Missing Information

- It turns out that this matrix is equal to

$$\boldsymbol{I} - \mathcal{I}_{\text{obs}}(\boldsymbol{\theta}^*)\mathcal{I}_{\text{com}}^{-1}(\boldsymbol{\theta}^*)$$

  - Here $\boldsymbol{I}$ is the identity matrix of length $p \times p$, where $p = \dim(\boldsymbol{\theta})$
- $\mathcal{I}_{\text{obs}}(\boldsymbol{\theta}^*)\mathcal{I}_{\text{com}}^{-1}(\boldsymbol{\theta}^*)$ is called the *fraction of missing information*
- With less missing data, $\mathcal{I}_{\text{obs}}(\boldsymbol{\theta}^*)\mathcal{I}_{\text{com}}^{-1}(\boldsymbol{\theta}^*)$ is "closer" to $\boldsymbol{I}$ and the rate of convergence improves
- Some components of $\boldsymbol{\theta}^{(t)}$ may have better convergence properties than others
  - Meng and Rubin [1994] give *componentwise* rates of convergence for the EM algorithm

References I

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Chuanhai Liu and Donald B Rubin. The ecme algorithm: a simple extension of em and ecm with faster monotone convergence. *Biometrika*, 81(4): 633–648, 1994.

Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

Xiao-Li Meng and Donald B Rubin. On the global and componentwise rates of convergence of the em algorithm. *Linear Algebra and its Applications*, 199:413–425, 1994.