Week 7: The EM-Algorithm

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2024-11-01

EM Algorithm - Recap

- $\mathbf{X}_{obs}$ are the **observed** random variables
- $\mathbf{X}_{miss}$ are the **missing** random variables
- $\ell_{comp}(\theta)$ is the **complete** log-likelihood of $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$
  - maximizing this to obtain MLE is supposed to be *simple*
  - θ denotes all the parameters, e.g., contains μ and Σ

Our task is to maximize $\ell_{obs}(\theta)$, the **observed** log-likelihood of $\mathbf{X}_{obs}$

EM Algorithm: Start from an initial estimate $\theta^{(0)}$ and for $l = 1, 2, ...$ iterate the following two steps until convergence:

- **E-step**: calculate $\mathbb{E}_{\widehat{\theta}^{(l-1)}}\big[\ell_{comp}(\theta)\big|\mathbf{X}_{obs} = \mathbf{x}_{obs}\big] =: Q(\theta, \theta^{(l-1)})$
- **M-step**: optimize $\underset{\theta}{\arg\max}\, Q\big(\theta, \theta^{(l-1)}\big) =: \theta^{(l)}$

Section 1

Some Properties of EM

Monotone Convergence

Proposition 1: $\ell_{obs}(\theta^{(l)}) \geq \ell_{obs}(\theta^{(l-1)})$

- a step of the EM algorithm will never decrease the objective value
- algorithms with this property are typically
  - numerically stable (good)
  - convergent under mild conditions (good)
- the algorithm is guaranteed to converge to a stationary point of the likelihood under a continuity condition on $Q(\cdot, \cdot)$; see Theorem 3.2 in McLachlan and Krishnan, 2007
  - convergence to a unique MLE requires unimodality of the likelihood (among other conditions)
  - prone to get stuck in local maxima (bad)

Monotone Convergence - Proof

The joint density for the complete data $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})^\top$ satisfies $f_\theta(\mathbf{X}) = f_\theta(\mathbf{X}_{miss}|\mathbf{X}_{obs}) f_\theta(\mathbf{X}_{obs})$ and hence

$$\ell_{comp}(\theta) = \log f_\theta(\mathbf{X}_{miss}|\mathbf{X}_{obs}) + \ell_{obs}(\theta)$$

Notice that $\ell_{obs}(\theta)$ does not depend on $\mathbf{X}_{miss}$ and hence we can condition on $\mathbf{X}_{obs}$ under any value of the parameter θ without really doing anything:

$$\begin{aligned}
\ell_{obs}(\theta) &= \mathbb{E}_{\theta^{(l-1)}}\left\{\ell_{comp}(\theta) - \log f_\theta(\mathbf{X}_{miss}|\mathbf{X}_{obs})\Big|X_{obs}\right\} \\
&= \underbrace{\mathbb{E}_{\theta^{(l-1)}}\left\{\ell_{comp}(\theta)\Big|X_{obs}\right\}}_{=Q\left(\theta,\theta^{(l-1)}\right)} - \underbrace{\mathbb{E}_{\theta^{(l-1)}}\left\{\log f_\theta(X_{miss}|X_{obs})\Big|X_{obs}\right\}}_{=:H\left(\theta,\theta^{(l-1)}\right)}
\end{aligned}$$

Thus, when we take $\widehat{\theta}^{(l)} = \arg\max_\theta Q\big(\theta, \widehat{\theta}^{(l-1)}\big)$, we only have to show that we have not increased $-H\big(\cdot, \theta^{(l-1)}\big)$

Monotone Convergence - Proof

Dividing and multiplying by $f_{\theta^{(l-1)}}(X_{miss}|X_{obs})$ and using the Jensen's inequality, we obtain just that:

$$H(\theta,\theta^{(l-1)}) = \mathbb{E}_{\theta^{(l-1)}}\left\{ \ln \frac{f_\theta(X_{miss}|X_{obs})}{f_{\theta^{(l-1)}}(X_{miss}|X_{obs})}\Big|X_{obs}\right\} + H\big(\theta^{(l-1)},\theta^{(l-1)}\big)$$

$$\leq \ln \underbrace{\mathbb{E}_{\theta^{(l-1)}}\left\{ \frac{f_\theta(X_{miss}|X_{obs})}{f_{\theta^{(l-1)}}(X_{miss}|X_{obs})}\Big|X_{obs}\right\}}_{=\int \frac{f_\theta(x_{miss}|X_{obs})}{f_{\theta^{(l-1)}}(x_{miss}|X_{obs})} f_{\theta^{(l-1)}}(x_{miss}|X_{obs})dx_{miss}=1} + H\big(\theta^{(l-1)},\theta^{(l-1)}\big)$$

and so indeed $H\big(\theta,\theta^{(l-1)}\big) \leq H\big(\theta^{(l-1)},\theta^{(l-1)}\big)$

Speed of Convergence: Definition

We have an iterative algorithm that is trying to find the maximum/minimum of a function and we want an estimate of how long it will take to reach that optimal value

For an iterative algorithm that converges to a solution $\Theta^\star$, if there is a real number γ and a constant integer k_0, such that for all $k > k_0$, we have

$$\left\|\Theta^{(k+1)} - \Theta^\star\right\| \leq q \left\|\Theta^{(k)} - \Theta^\star\right\|^\gamma$$

with q being a positive constant independent of k, then we say that the algorithm has a convergence rate of order γ. An algorithm has

- first-order or linear convergence if $\gamma = 1$ and $q \in (0, 1)$ (sublinear if $q = 1$)
- superlinear convergence if $1 < \gamma < 2$ (quasi-Newton, method of scoring)
- second-order or quadratic convergence if $\gamma = 2$ (Newton)

Speed of Convergence for EM

Consider the iteration mapping $M : \theta^{(l-1)} \mapsto \theta^{(l)}$, assumed continuous

- if $\theta^{(l)} \to \theta^\star$ as $l \to \infty$, then it must be a fixed point: $M(\theta^\star) = \theta^\star$
- in the neighborhood of $\theta^\star$, a 1st order Taylor expansion:

$$\theta^{(l+1)} = M(\theta^{(l)}) \approx \theta^\star + \left.\frac{\partial M(\theta)}{\partial \theta^\top}\right|_{\theta=\theta^\star} (\theta^{(l)} - \theta^\star)$$

  yields

$$\theta^{(l+1)} - \theta^\star \approx \mathbf{J}(\theta^\star)\, (\theta^{(l)} - \theta^\star),$$

  where $\mathbf{J}(\theta^\star)$ is the Jacobian matrix and measures the rate of convergence
- Smaller $\|\mathbf{J}(\theta^\star)\| = \lim \|\theta^{(l+1)} - \theta^{(l)}\| / \|\theta^{(l)} - \theta^{(l-1)}\|$ means faster global conv.
- Rate is linear: $\|\theta^{(l)} - \theta^\star\| \approx \|\mathbf{J}(\theta^\star)\|^l \, \|\theta^{(0)} - \theta^\star\|$
- If $\|\mathbf{J}(\theta^\star)\| < 1$, then M is a contraction and we may hope for convergence

Speed of Convergence for EM

Consider the iteration mapping $M : \theta^{(l-1)} \mapsto \theta^{(l)}$, assumed continuous

- if $\theta^{(l)} \to \theta^\star$ as $l \to \infty$, then it must be a fixed point: $M(\theta^\star) = \theta^\star$
- in the neighborhood of $\theta^\star$, a 1st order Taylor expansion:

$$\theta^{(l+1)} = M(\theta^{(l)}) \approx \theta^\star + \left.\frac{\partial M(\theta)}{\partial \theta^\top}\right|_{\theta=\theta^\star} (\theta^{(l)} - \theta^\star)$$

  yields

$$\theta^{(l+1)} - \theta^\star \approx \mathbf{J}(\theta^\star)\ (\theta^{(l)} - \theta^\star),$$

  where $\mathbf{J}(\theta^\star)$ is the Jacobian matrix and measures the rate of convergence
- Smaller $\|\mathbf{J}(\theta^\star)\| = \lim \|\theta^{(l+1)} - \theta^{(l)}\| / \|\theta^{(l)} - \theta^{(l-1)}\|$ means faster global conv.
- Rate is linear: $\|\theta^{(l)} - \theta^\star\| \approx \|\mathbf{J}(\theta^\star)\|^l\ \|\theta^{(0)} - \theta^\star\|$
- If $\|\mathbf{J}(\theta^\star)\| < 1$, then M is a contraction and we may hope for convergence

It can be shown that:

$$\mathbf{J}(\theta^\star) = \mathbf{J}_{comp}^{-1}(\theta^\star)\ \mathbf{J}_{miss}(\theta^\star),$$

where $\mathbf{J}_{comp}$ and $\mathbf{J}_{miss}$ are Fisher information of the complete resp. missing data

$\Rightarrow$ the bigger the proportion of missing information, the slower the convergence

Exponential Families

Let the density of the complete data be from the exponential family, i.e.,

$$f_X(\mathbf{x}) = \exp\left\{\eta(\theta)^\top \mathbf{T}(\mathbf{x}) - g(\theta)\right\} h(\mathbf{x})$$

where

- $\theta \in \Theta \subset \mathbb{R}^p$
- $\mathbf{T}(\mathbf{x}) = \big(T_1(\mathbf{x}), \dots, T_p(\mathbf{x})\big)^\top$ is the *sufficient statistic* for θ
- $\eta : \mathbb{R}^p \to \mathbb{R}^p$, $g : \mathbb{R}^p \to R$ and $h : \mathbb{R}^d \to \mathbb{R}$

Assuming $\eta(\theta) = \theta$, i.e., θ is the canonical parameter, we have

$$\ell_{comp}(\theta) = \sum_{n=1}^{N} \theta^\top \mathbf{T}(\mathbf{X}_n) + \ln h(X_n) - Ng(\theta)$$

and

$$Q(\theta, \theta^{(l-1)}) = \sum_{n=1}^{N} \theta^\top \mathbf{t}_n^{(l)} + \mathbb{E}_{\theta^{(l-1)}}\big[\ln h(X_n)\big|\mathbf{X}_{obs}\big] - Ng(\theta),$$

where $t_n^{(l)} = \mathbb{E}_{\theta^{(l-1)}}\big[T(\mathbf{X}_n)\big|\mathbf{X}_{obs}\big]$

Exponential Families

- It is straightforward that for the E-step we will only need to compute the conditional expectations of the complete-data sufficient statistics

$$\mathbb{E}_{\theta^{(l-1)}}\big[T_i(\mathbf{X})\big|\mathbf{X}_{obs}\big], \quad i = 1, \dots, p$$

- The M-step is equivalent to finding the expressions for the complete-data ML estimates of θ and replacing the complete-data sufficient statistics in these expressions with their conditional expectations computed in the E step

Note: This applies, e.g., to Example 3 from Week 6

Section 2

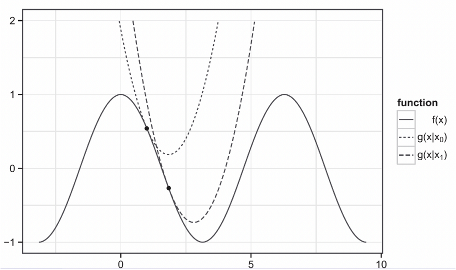
MM Algorithms

MM Algorithms

Definition: A function $g(\mathbf{x} \mid \mathbf{x}^{(l)})$ is said to **majorize** a function $f : \mathbb{R}^p \to \mathbb{R}$ at $\mathbf{x}^{(l)}$ provided

- $f(\mathbf{x}) \leq g(\mathbf{x}|\mathbf{x}^{(l)}), \quad \forall\, \mathbf{x}$
- $f(\mathbf{x}^{(l)}) = g(\mathbf{x}^{(l)}|\mathbf{x}^{(l)})$

In other words, the surface $\mathbf{x} \mapsto g(\mathbf{x}|\mathbf{x}^{(l)})$ is above the surface $f(\mathbf{x})$, and it is touching it at $\mathbf{x}^{(l)}$



MM Algorithms

Assume our goal is to minimize a function $f : \mathbb{R}^p \rightarrow \mathbb{R}$

The basic idea of the MM algorithm is to start from an initial guess $\mathbf{x}^{(0)}$ and for $l = 1, 2, \ldots$ iterate between the following steps until convergence:

- **Majorization step**: construct $g(\mathbf{x}|\mathbf{x}^{(l-1)})$, i.e., construct a majorizing function to f at $\mathbf{x}^{(l-1)}$
- **Minimization step**: set $\mathbf{x}^{(l)} = \arg\min_{\mathbf{x}} g(\mathbf{x}|\mathbf{x}^{(l-1)})$, i.e., minimize the majorizing function

$\rightarrow$ MM stands for "Majorization-Minimization" or "Minorization-Maximization"

Monotone convergence property is trivially guaranteed by construction:

$$f(\mathbf{x}^{(l)}) \leq g(\mathbf{x}^{(l)}|\mathbf{x}^{(l-1)}) \leq g(\mathbf{x}^{(l-1)}|\mathbf{x}^{(l-1)}) = f(\mathbf{x}^{(l-1)})$$

E-step Minorizes

With extra minus sign, the EM is:

$$\begin{aligned} &\textbf{E-step:} \quad Q(\theta|\theta^{(l-1)}) := \mathbb{E}_{\theta^{(l-1)}}\big[-\ell_{comp}(\theta)\big|X_{obs}\big] \\ &\textbf{M-step:} \qquad\quad \theta^{(l)} := \arg\min_{\theta} Q(\theta|\theta^{(l-1)}) \end{aligned}$$

From the proof of Proposition 1 above, we have (with the extra sign)

$$-\ell_{obs}(\theta) = -Q(\theta|\theta^{(l-1)}) + H(\theta, \theta^{(l-1)})$$

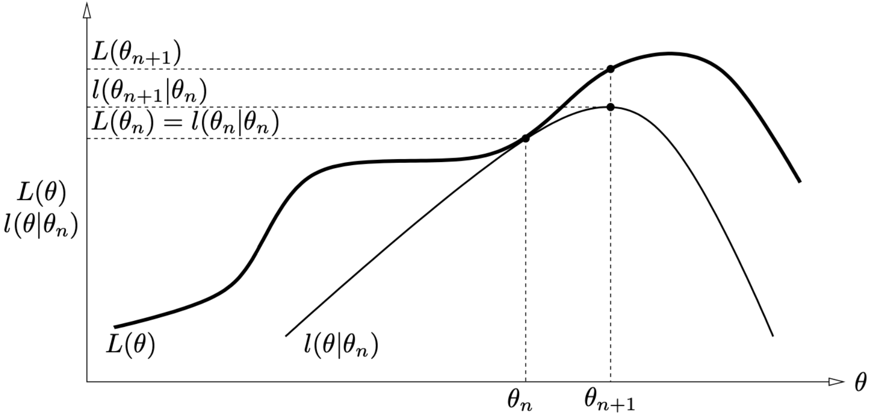
and since $H(\theta, \theta^{(l-1)}) \leq H(\theta^{(l-1)}, \theta^{(l-1)})$, we obtain

$$-\ell_{obs}(\theta) \leq -Q(\theta|\theta^{(l-1)}) + H(\theta^{(l-1)}, \theta^{(l-1)}) =: \widetilde{Q}(\theta|\theta^{(l-1)})$$

with equality at $\theta = \theta^{(l-1)}$

- $\widetilde{Q}(\theta|\theta^{(l-1)})$ is majorizing $-\ell_{obs}(\theta)$ at $\theta = \theta^{(l-1)}$
- $H(\theta^{(l-1)}, \theta^{(l-1)})$ is a constant (w.r.t. θ)

Graphical interpretation Revisited

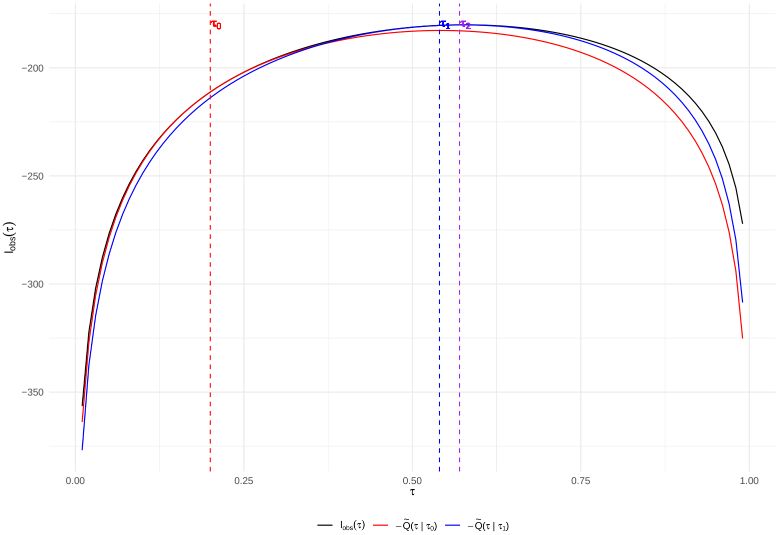


$$\ell(\theta \mid \theta_n) = -\tilde{Q}(\theta \mid \theta_n) = Q(\theta|\theta^{(n)}) - Q(\theta^{(n)}|\theta^{(n)}) + \ell_{obs}(\theta^{(n)}) \leq \ell_{obs}(\theta) = L(\theta)$$

Example 2 (Week 6) Revisited

```
rmixnorm <- function(N, tau, mu1=3, mu2=0, sigma1=0.5, sigma2=1){
  ind <- I(runif(N) > tau)
  X <- rep(0,N)
  X[ind] <- rnorm(sum(ind), mu1, sigma1)
  X[!ind] <- rnorm(sum(!ind), mu2, sigma2)
  return(X)
}
dmixnorm <- function(x, tau, mu1=3, mu2=0, sigma1=0.5, sigma2=1){
  y <- (1-tau)*dnorm(x,mu1,sigma1) + tau*dnorm(x,mu2,sigma2)
  return(y)
}
ell_obs <- function(X, tau, mu1=3, mu2=0, sigma1=0.5, sigma2=1){
  return(sum(log(dmixnorm(X, tau, mu1, mu2, sigma1, sigma2))))
}
Q <- function(t, tl){
  gammas <- dnorm(X)*tl/dmixnorm(X, tl)
  qs <- dnorm(X,3,0.5)^(1-gammas)*dnorm(X)^gammas*t^gammas*(1-t)^(1-gammas)
  return(sum(log(qs)))
}
```

Two Steps Visualized



MM example: Finding a (sample) median

Consider the sequence of observations $x_1, \dots, x_N$. The sample median θ minimizes the non-differentiable criterion

$$f(\theta) = \sum_{n=1}^{N} |x_n - \theta|$$

The quadratic function

$$h_n\left(\theta \mid \theta^l\right) = \frac{1}{2}\frac{\left(x_n - \theta\right)^2}{\left|x_n - \theta^l\right|} + \frac{1}{2}\left|x_n - \theta^l\right|$$

majorizes $|x_n - \theta|$ at $\theta^l \Rightarrow g\left(\theta \mid \theta^l\right) = \sum_{n=1}^{N} h_n\left(\theta \mid \theta^l\right)$ majorizes $f(\theta)$

The minimum of $g\left(\theta \mid \theta^l\right)$ occurs at $\theta^{l+1} = (\sum_{n=1}^{N} w_n^l x_n)/(\sum_{n=1}^{N} w_n^l)$, for $w_n^l = \left|x_n - \theta^l\right|^{-1}$

$\rightarrow$ generalizes to L_1 regression and quantile regression

MM Convergence

Theorem. (Lange, 2013, Proposition 12.4.4)
Suppose that all stationary points of $f(\mathbf{x})$ are isolated and that the *differentiability*, *coerciveness*, and *convexity* assumptions are true. Then any sequence that iterates $\mathbf{x}^{(l)} = M(\mathbf{x}^{(l-1)})$, generated by the iteration map $M(\cdot)$ of the MM algorithm, possesses a limit, and that limit is a stationary point of $f(\mathbf{x})$. If $f(\mathbf{x})$ is strictly convex, then $\lim_{l\to\infty} \mathbf{x}^{(l)}$ is the minimum point.

- *differentiability* - conditions on majorizations guaranteeing differentiability of the iteration map M
- *coerciveness* - upper level sets of f $\{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ are compact (ensures that local maxima do not occur on the boundary)
- *convexity* - just technical! Without it, we would say that all limit points (which however might not exist without convexity) are stationary points and MM converges to one of them

MM Remarks

- MM algorithms can linearize an optimization problem (mixture of Gaussians)
- MM algorithms can turn a non-differentiable problem into a smooth problem
- The rate of convergence depends on how well the majorizer/minorizer $g(\mathbf{x} \mid \mathbf{x}^{(l)})$ approximates the target $f(\mathbf{x})$
- There exist methods for accelerating the convergence of MM and EM algorithms (e.g., Aitken's method); see Zhou et al. (2009) and Chapter 4 in McLachlan and Krishnan (2008)

Concluding EM Remarks

- EM is just MM with majorization achieved by Jensen's inequality
- due to the monotone convergence property of all MM algorithms, EM
  - is numerically stable
  - typically converges
  - but can get stuck in a local minimum/maximum

How to choose starting parameters in mixture of Gaussian?

Hastie and Tibshirani (Elements of Statistical Learning, pg. 293) recommend constructing initial guesses as follows:

- For $\hat{\mu_1}$ and $\hat{\mu_2}$, randomly select two y_i values
- For $\hat{\Sigma}_1^2$ and $\hat{\Sigma}_2^2$, set both equal to the overall sample variance $\sum_{i=1}^{N}\left(y_i - \bar{y}\right)\left(y_i - \bar{y}\right)^T / N$
- For $\hat{\pi}$, begin at 0.50

In practice, the EM algorithm is often run using several different combinations of starting parameter estimates $\Rightarrow$ prevents relying on one set of starting parameters that may get stuck in a local max

Concluding EM Remarks

- EM computational costs per iteration are typically favorable (simple steps), but
- convergence relatively slow (many steps)
  - linear at the neighborhood of the limit
  - in practice monitored by looking at $\|\mathbf{x}^{(l)} - \mathbf{x}^{(l-1)}\|$ and $|f(\mathbf{x}^{(l)}) - f(\mathbf{x}^{(l-1)})|$
- the M-step may not have a closed form solution, but is typically much simpler than the original problem
  - if inner iteration for the M-step, early stopping is often desirable
  - ex.: logistic regression with missing covariates (M-step solved by IRLS)

References

- Lange, K. (2013). *Optimization*. 2nd Edition.
- Lange, K. (2016). *MM optimization algorithms*.
- McLachlan, G.J., & Krishan, T. (2008). *The EM algorithm and extensions*.

Main Project

Go to Main project for details