Week 7: The EM-Algorithm

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2023-11-03

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 $l = 1, 2, \ldots$ iterate the following two steps until convergence:

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

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; 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 minima (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:

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

Thus, when we take $\widehat{\theta}^{(l)} = \arg\max_\theta Q(\theta, \widehat{\theta}^{(l-1)})$, 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:

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

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

Speed of Convergence

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

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

  where $\mathbf{J}(\theta^\star)$ is the Jacobian matrix and measures the rate of convergence
- Smaller $\|\mathbf{J}(\theta^\star)\| = \lim \|\theta^{(l+1)} - \theta^{(l)}\| / \|\theta^{(l)} - \theta^{(l-1)}\|$ mean faster 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

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

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

  where $\mathbf{J}(\theta^\star)$ is the Jacobian matrix and measures the rate of convergence
- Smaller $\|\mathbf{J}(\theta^\star)\| = \lim \|\theta^{(l+1)} - \theta^{(l)}\| / \|\theta^{(l)} - \theta^{(l-1)}\|$ mean faster 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 $h : \mathbb{R}^p \to \mathbb{R}$

It is straightforward that for the E-step we will only need to calculate the following expectations

$$\mathbb{E}_{\theta^{(l-1)}}\left[T_i(\mathbf{X}) \middle| \mathbf{X}_{obs}\right]$$

and plug them into the likelihood instead of the complete data sufficient statistic

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

Section 2

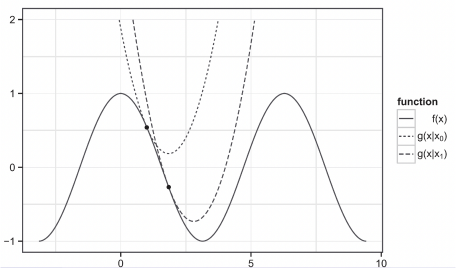
MM Algorithms

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 $f : \mathbb{R}^p \to \mathbb{R}$

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

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

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

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

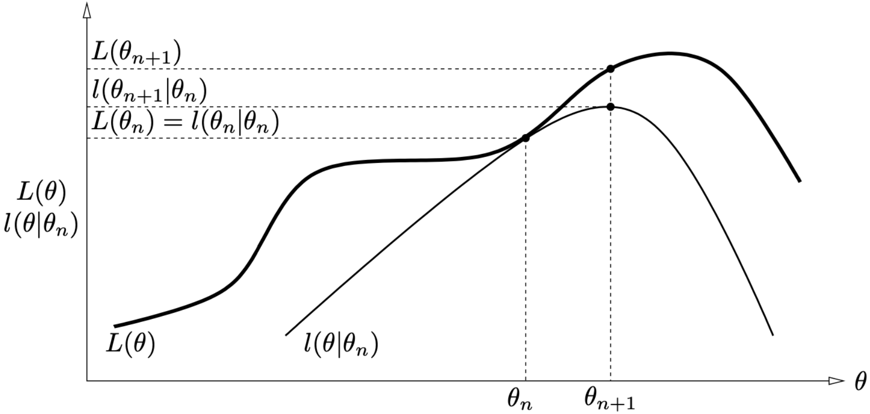
and since $H(\theta, \theta^{(l-1)}) \leq H(\theta^{(l-1)}, \theta^{(l-1)})$, we obtain

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

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



$\ell(\theta \mid \theta_n) = -\tilde{Q}(\theta \mid \theta_n) = Q(\theta|\theta^{(l-1)}) - Q(\theta^{(l-1)}|\theta^{(l-1)}) + \ell_{obs}(\theta^{(l-1)}) \leq \ell_{obs}(\theta) = L(\theta)$

Example 2 (Week 6) Revisited

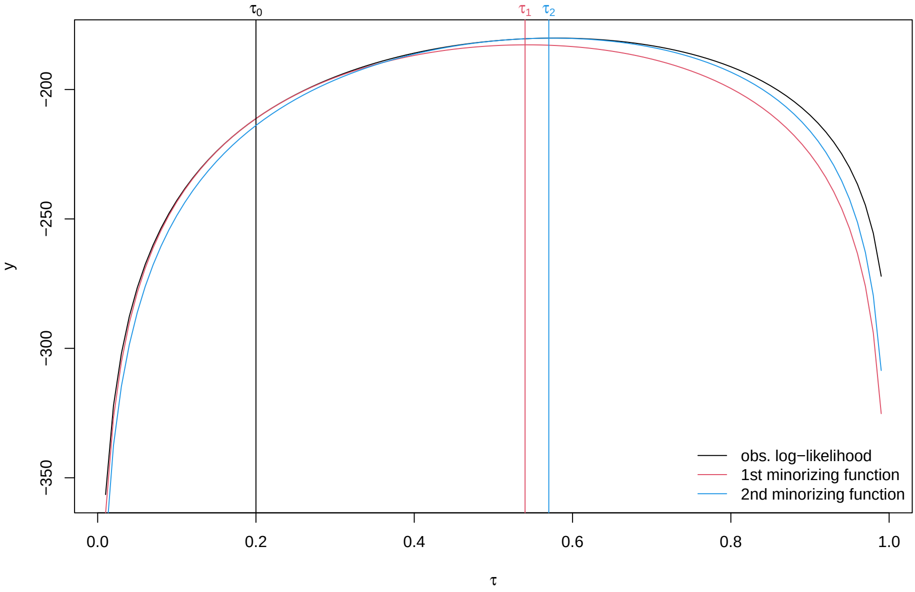
```
rmixnorm <- function(N, tau, mu1=3, mu2=0, sigma1=0.5, sigma2=1){
  ind <- I(runif(N) > tau)
  X <- rep(0,N)
  X[ind] <- rnorm(sum(ind), mu1, sigma1)
  X[!ind] <- rnorm(sum(!ind), mu2, sigma2)
  return(X)
}
dmixnorm <- function(x, tau, mu1=3, mu2=0, sigma1=0.5, sigma2=1){
  y <- (1-tau)*dnorm(x,mu1,sigma1)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

```
N <- 100
tau <- 0.6
set.seed(517)
X <- rmixnorm(N, tau)
t <- seq(0.01,0.99,by=0.01)
y <- rep(0,99)
for(i in 1:99) y[i] <- ell_obs(X,t[i])

plot(t,y,type="l", xlab=expression(tau))
y0 <- rep(0,99)
for(i in 1:99) y0[i] <- Q(t[i],t[20])
points(t,y0-y0[20]+y[20],type="l",lty=1, col=2)
abline(v=t[20])
mtext(expression(tau[0]), at = t[20], side = 3)
ind <- which(y0==max(y0))
y1 <- rep(0,99)
for(i in 1:99) y1[i] <- Q(t[i],t[ind])
points(t,y1-y1[ind]+y[ind],type="l",lty=1,col=4,cex=1.5)
abline(v=t[ind],lty=1, col=2)
mtext(expression(tau[1]), at = t[ind], side = 3, col=2)
indnew <- which(y1==max(y1)); abline(v=t[indnew],lty=1,col=4,cex=1.5)
mtext(expression(tau[2]), at = t[indnew], side = 3, col=4)
legend("bottomright",
       c("obs. log-likelihood", "1st minorizing function", "2nd minorizing function"),
       col = c(1, 2, 4), lty = 1, bty = "n")
```

Two Steps Visualized



MM Convergence

Theorem. (Lange, 2013, Proposition 12.4.4)
Suppose that all stationary points of $f(\mathbf{x})$ are isolated and that the stated *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 are compact ($f(x) \to -\infty$ for $\|x\| \to \infty$)
- *convexity* - just technical! Without it, we would say that all limit points (which however might not exist without convexity) are stationary points

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
- EM computational costs per iteration are typically favorable
- convergence relatively slow
  - 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. (2007). *The EM algorithm and extensions*.

Main Project

Go to Main project for details