Week 8: The EM-Algorithm

MATH-517 Statistical Computation and Visualization

Tomas Masak

November 11th 2022

EM Algorithm - Recap

- $\boldsymbol{X}_{obs}$ are the **observed** random variables
- $\boldsymbol{X}_{miss}$ are the **missing** random variables
- $\ell_{comp}(\theta)$ is the **complete** log-likelihood of $\boldsymbol{X} = (\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss})$
  - 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 $\boldsymbol{X}_{obs}$.

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 $\mathbb{E}_{\widehat{\theta}^{(l-1)}}\big[\ell_{comp}(\theta)\big|\boldsymbol{X}_{obs} = \mathbf{x}_{obs}\big] =: Q\big(\theta, \theta^{(l-1)}\big)$
- **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)
  - prone to get stuck in local minima (bad)

Monotone Convergence - Proof

The joint density for the complete data X satisfies $f_\theta(X) = f_\theta(X_{miss}|X_{obs}) f_\theta(X_{obs})$ and hence

$$\ell_{comp}(\theta) = \log f_\theta(X_{miss}|X_{obs}) + \ell_{obs}(\theta).$$

Notice that $\ell_{obs}(\theta) = \ell_{comp}(\theta) - \log f_\theta(X_{miss}|X_{obs})$ does not depend on X_{miss} ($\ell_{obs}(\theta)$ clearly does not) and hence we can condition on X_{obs} under any value of the parameter θ without really doing anything:

$$\ell_{obs}(\theta) = \underbrace{\mathbb{E}_{\theta^{(l-1)}}\left[\ell_{comp}(\theta)\middle|X_{obs}\right]}_{=Q\left(\theta,\theta^{(l-1)}\right)} - \underbrace{\mathbb{E}_{\theta^{(l-1)}}\left[\log f_\theta(X_{miss}|X_{obs})\middle|X_{obs}\right]}_{=:H\left(\theta,\theta^{(l-1)}\right)}$$

And so when we take $\widehat{\theta}^{(l)} = \arg\max_\theta Q(\theta, \widehat{\theta}^{(l-1)})$, we only have to show 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:

$$H(\theta, \theta^{(l-1)}) = \mathbb{E}_{\theta^{(l-1)}}\left[\log \frac{f_\theta(X_{miss}|X_{obs})}{f_{\theta^{(l-1)}}(X_{miss}|X_{obs})}\middle| X_{obs}\right] + H(\theta^{(l-1)}, \theta^{(l-1)})$$

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

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

Speed of Convergence

$$M : \theta^{(l-1)} \mapsto \theta^{(l)}$$

- 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 $\theta^\star$, by Taylor:

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

  where $\mathbf{J}(\theta^\star)$ is the Jacobian
- If $\|\mathbf{J}(\theta^\star)\| < 1$, then M is a contraction and we may hope for convergence, with smaller $\|\mathbf{J}(\theta^\star)\|$ corresponding to a faster convergence speed
  - rate is linear: $\|\theta^{(l)} - \theta^\star\| \approx \|\mathbf{J}(\theta^\star)\|^l \, \|\theta^{(0)} - \theta^\star\|$

Speed of Convergence

$$M : \theta^{(l-1)} \mapsto \theta^{(l)}$$

- 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 $\theta^\star$, by Taylor:

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

  where $\mathbf{J}(\theta^\star)$ is the Jacobian
- If $\|\mathbf{J}(\theta^\star)\| < 1$, then M is a contraction and we may hope for convergence, with smaller $\|\mathbf{J}(\theta^\star)\|$ corresponding to a faster convergence speed
  - rate is linear: $\|\theta^{(l)} - \theta^\star\| \approx \|\mathbf{J}(\theta^\star)\|^l \, \|\theta^{(0)} - \theta^\star\|$

It can be shown:

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

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

Exponential Families

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

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

where

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

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

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

and plug them into the likelihood.

Note: While this applies e.g. to Example 3 from Week 7, it is not that useful apart from giving us confidence.

Section 2

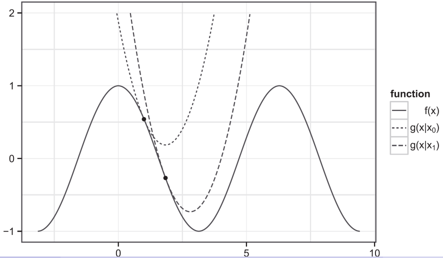
MM Algorithms

MM Algorithms

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

$$
\begin{aligned}
f(\mathbf{x}) \leq g(\mathbf{x}|\mathbf{x}^{(l)}), &\qquad \forall\, \mathbf{x}, \\
f(\mathbf{x}^{(l)}) = g(\mathbf{x}^{(l)}|\mathbf{x}^{(l)}). &
\end{aligned}
$$

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

Monotone convergence property trivially by the construction:

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

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 $\theta = \theta^{(-l)}$
- $H(\theta^{(l-1)}, \theta^{(l-1)})$ is a constant (w.r.t. θ)

Example 2 (Week 7) Revisited

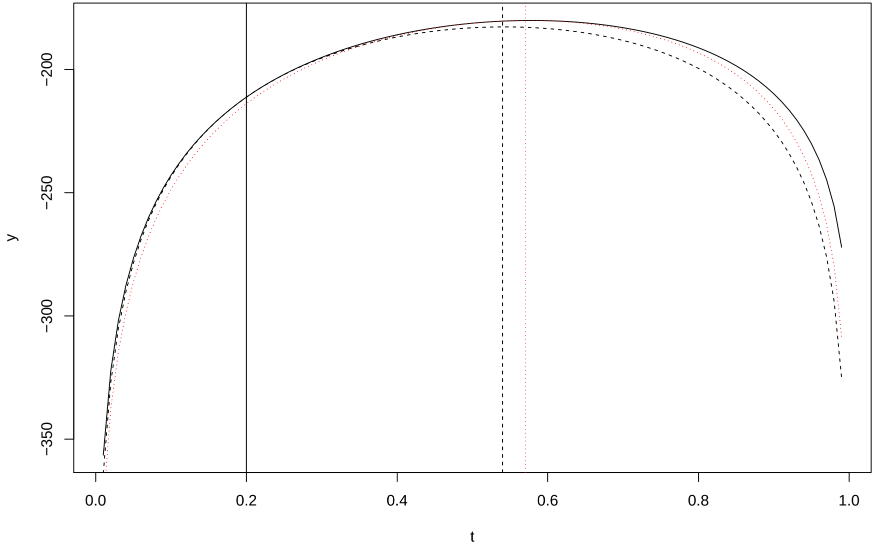
```
rmixnorm <- function(N, tau, mu1=3, mu2=0, sigma1=0.5, sigma2=1){
  ind <- I(runif(N) > tau)
  X <- rep(0,N)
  X[ind] <- rnorm(sum(ind), mu1, sigma1)
  X[!ind] <- rnorm(sum(!ind), mu2, sigma2)
  return(X)
}
dmixnorm <- function(x, tau, mu1=3, mu2=0, sigma1=0.5, sigma2=1){
  y <- (1-tau)*dnorm(x,mu1,sigma1) + 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")
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=2)
abline(v=t[20])
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=3,col="red",cex=1.5)
abline(v=t[ind],lty=2)
indnew <- which(y1==max(y1))
abline(v=t[indnew],lty=3,col="red",cex=1.5)
```

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 of 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 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
  - can get stuck in a local minimum
- 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)

Example 2 (Week 7) Revisited

```r
mixtureEM <- function(X, mu1, mu2, sigma1, sigma2, tau){
  ...
  while("not converged"){
    ...
    print(ell_obs(X, mu1, mu2, sigma1, sigma2, tau))
  }
  return(list(mu1, mu2, sigma1, sigma2, tau))
}

N <- 100
mu1 <- 3; mu2 <- 0
sigma1 <- 0.5; sigma2 <- 1
tau <- 0.6

X <- rmixnorm(N, mu1, mu2, sigma1, sigma2, tau)
unlist(mixtureEM(X, 1, 1, 1, 1, 0.5))
```

```
[1] -189.5242
[1] -189.5242
[1] 1.356270 1.356270 1.610112 1.610112 0.500000
```

References

- Lange, K. (2013). *Optimization*. 2nd Edition.
- Lange, K. (2016). *MM optimization algorithms*.
- McLachlan, G.J., & Krishan, T. (2007). *The EM algorithm and extensions*.

Project Ideas

- Cross-validation for PCA
  - A simulation study to compare the advantages of EM compared to what we did last week.
  - Two more approaches to CV for PCA to be considered and compared.
- Comparison of local regression implementations in different R packages.
  - Wickham (2011) examines several packages for KDE in R. Not only there are huge differences in terms of speed, but some of the packages are even inconsistent! Makes one wonder what is the situation with local regression.
- Various simulation studies on numerous bandwidth selection strategies in KDE or local linear regression, potentially exploring available software.
  - e.g. direct comparison of several non-standard approaches via a simulation study

Further Ideas

Different simulation studies, e.g.

- Comparison of variable selectors in regression.
  - Hastie et al. (2020) have some surprising results in their simulation study, but one important method (adaptive lasso) is omitted. Try to recreate the study with adaptive lasso included.
- The simulation study above can be tackled from the prediction perspective as well.
  - How PCA, PLS (partial least squares) and ridge regressions compare against each other w.r.t. prediction?

Diving into one of the course topics.

- e.g. MM algorithms

Data analysis.

- e.g. if missing data in your small project $\Rightarrow$ EM

Consult your choices!