Gibbs sampling (an MCMC method) and relations to EM Lectures – Outline

Part 1 (Feb. 20) Gibbs

- the algorithm
- a bivariate example
- an elementary convergence proof for a (discrete) bivariate case
- more than two variables
- a counter example.

Part 2 (Feb. 25) EM – again

- **EM as a maximization/maximization method**
    - **Gibbs as a variation of Generalized EM with an example (for HW #2)**
- **A counterexample for EM**

EM as a maximization/maximization method.

Recall:

$\mathbf{L}(\theta\,;\boldsymbol{x})$ is the likelihood function for θ with respect to the incomplete data $\boldsymbol{x}$.

$\mathbf{L}(\theta\,;(\boldsymbol{x},\boldsymbol{z}))$ is the likelihood for θ with respect to the complete data $(\boldsymbol{x},\boldsymbol{z})$.

And $\mathbf{L}(\theta\,;\boldsymbol{z}\mid\boldsymbol{x})$ is a *conditional likelihood* for θ with respect to $\boldsymbol{z}$, given $\boldsymbol{x}$;

which is based on $\boldsymbol{h}(\boldsymbol{z}\mid\boldsymbol{x},\theta)$: the conditional density for the data $\boldsymbol{z}$, given $(\boldsymbol{x},\theta)$.

Then as $$\boldsymbol{f}(\boldsymbol{X}\mid\theta) = \boldsymbol{f}(\boldsymbol{X},\boldsymbol{Z}\mid\theta)\ /\ \boldsymbol{h}(\boldsymbol{Z}\mid\boldsymbol{x},\theta)$$

we have $$\boldsymbol{log}\ \mathrm{L}(\theta\,;\boldsymbol{x}) = \boldsymbol{log}\ \mathrm{L}(\theta\,;(\boldsymbol{x},\boldsymbol{z})) - \boldsymbol{log}\ \mathrm{L}(\theta\,;\boldsymbol{z}\mid\boldsymbol{x}) \qquad (*)$$

As below, we use the *EM* algorithm to compute the *mle*

$$\hat{\theta} \;=\; \boldsymbol{argmax}_{\Theta}\ \mathrm{L}(\theta\,;\boldsymbol{x})$$

With $\hat{\theta}_0$ an arbitrary choice, define

(***E**-step)* $\qquad \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_0) \;=\; \int_Z [\boldsymbol{log}\ \mathrm{L}(\theta\, ;\, \boldsymbol{x}, \boldsymbol{z})]\ \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_0)\ dz$

and

$$\boldsymbol{H}(\theta \mid \boldsymbol{x}, \hat{\theta}_0) \;=\; \int_Z [\boldsymbol{log}\ \mathrm{L}(\theta\, ;\, \boldsymbol{z} \mid \boldsymbol{x})]\ \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_0)\ d\boldsymbol{z}.$$

then $\qquad \boldsymbol{log}\ \mathrm{L}(\theta\, ;\, \boldsymbol{x}) \;=\; \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0) - \boldsymbol{H}(\theta \mid \boldsymbol{x}, \theta_0),$

as we have integrated-out $\boldsymbol{z}$ from (*) using the conditional density $\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_0)$.

The ***EM algorithm*** is an iteration of

(1) the ***E***-step: determine the integral $\boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_j)$,

(2) the ***M***-step: define $\hat{\theta}_{j+1}$ as $\boldsymbol{argmax}_{\Theta}\ \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_j)$.

Continue until there is convergence of the $\hat{\theta}_j$.

Now, for a *Generalized EM* algorithm.

Let be $\boldsymbol{P}(\boldsymbol{Z})$ any distribution over the augmented data $\mathbf{Z}$, with density $\boldsymbol{p}(\boldsymbol{z})$
Define the function $\boldsymbol{F}$ by**:**

$$\boldsymbol{F}(\theta, \boldsymbol{P}(\boldsymbol{Z})) = \int_Z [\boldsymbol{log}\,\mathbf{L}(\theta; \boldsymbol{x}, \boldsymbol{z})]\, \boldsymbol{p}(\boldsymbol{z})dz - \int_Z \boldsymbol{log}\,\boldsymbol{p}(\boldsymbol{z})\, \boldsymbol{p}(\boldsymbol{z})dz$$

$$= \mathbf{E}_{\boldsymbol{P}}[\boldsymbol{log}\,\mathbf{L}(\theta; \boldsymbol{x}, \boldsymbol{z})] - \mathbf{E}_{\boldsymbol{P}}[\,\boldsymbol{log}\,\boldsymbol{p}(\boldsymbol{z})]$$

When $\boldsymbol{p}(\boldsymbol{Z}) = \boldsymbol{h}(\boldsymbol{Z} \mid \boldsymbol{x}, \hat{\theta}_0)$ from above, then $\boldsymbol{F}(\theta, \boldsymbol{P}(\boldsymbol{Z})) = \boldsymbol{log}\,\mathbf{L}(\theta\,; \boldsymbol{x})$.

Claim: For a fixed (arbitrary) value $\theta = \hat{\theta}_0$, $\boldsymbol{F}(\hat{\theta}_0, \boldsymbol{P}(\boldsymbol{Z}))$ is maximized over distributions $\boldsymbol{P}(\boldsymbol{Z})$ by choosing $\boldsymbol{p}(\boldsymbol{Z}) = \boldsymbol{h}(\boldsymbol{Z} \mid \boldsymbol{x}, \hat{\theta}_0)$.

Thus, the *EM* algorithm is a sequence of ***M***-***M*** steps: the old ***E***-step now is a max over the second term in $\boldsymbol{F}(\hat{\theta}_0, \boldsymbol{P}(\boldsymbol{Z}))$, given the first term. The second step remains (as in *EM*) a max over θ for a fixed second term, which does not involve θ

Suppose that the augmented data $\boldsymbol{Z}$ are multidimensional.

Consider the *GEM* approach and, instead of maximizing the choice of $\boldsymbol{P}(\boldsymbol{Z})$ over all of the augmented data – instead of the old *E*-step – instead maximize over only *one* coordinate of $\boldsymbol{Z}$ at a time, alternating with the (old) ***M***-step.

This gives us the following link with the Gibbs algorithm: Instead of maximizing at each of these two steps, use the conditional distributions, we sample from them!

In HW #2, you will work out this parallel analysis between the *EM* and Gibbs algorithms for the calculation of the posterior distribution in the (k = 2) case of a *Mixture of Gaussians* problem.

An *EM* "*counterexample*":

We are testing failure times on a new variety of hard disk.
Based on an *ECE theory* of these disks, the failure times follow a
Uniform $\boldsymbol{U}(0, \theta]$ distribution, $\theta > 0$.

We select at random $m + n$ disks, having a common θ for failure
We select n of these (at random) and test them until failure.

These n disks run as *iid* $\boldsymbol{U}(0, \theta]$ quantities until they fail.
The lab records the data of their exact failure times: $\boldsymbol{y} = (y_1, \ldots, y_n)$.

We know (from HW #1) that

$$\hat{y} = \boldsymbol{max}\,(y_1, \ldots, y_n)$$

is both *sufficient* and is the *mle* for θ, w.r.t. the data $\boldsymbol{y}$.

We conduct a different experiment with the remaining m disks.

We start them at a common time $t_0 = 0$. At time $t > 0$, chosen as an ancillary quantity w.r.t. θ, we halt our m-trials and observe only which of the m-many disks are still running.

Thus our observed data from the second experiment are only the m indicators, $\boldsymbol{x} = (x_1, \ldots, x_m)$
where $x_i = 1$, or $x_i = 0$ as disk i is, or is not still running after t units time.

In what follows, assume that *at least* one of these m-disks is still running. So, given $\boldsymbol{x}$, we know that $\theta \geq t$.

Our goal is to calculate the *mle* $\hat{\theta}$

$$= \boldsymbol{argmax}_{\Theta}\, \mathbf{L}(\theta\ ;\ t,\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{argmax}_{\Theta}\, log\, \mathbf{L}(\theta\ ;\ t,\boldsymbol{x},\boldsymbol{y}) \quad \text{(as } log \text{ is monotone)}$$

The data $\boldsymbol{x}$ data are *incomplete* relative to data $\boldsymbol{y}$. We don't know the failure times for the m observed disks, though we have one-sided censoring for each.

That is, for $x_i = 0$, the i^{th} disk has already failed though we don't know its value. For $x_i = 1$, we may imagine, instead of halting the trial, letting the i^{th} disk continue to run until it would fail.

Denote these missing data correspond to $\boldsymbol{x}$ by $\boldsymbol{z} = (z_1, \ldots., z_m)$.
Thus, we have that $z_i > (\leq)\ t$ as $x_i = 1$ $(x_i = 0)$.
Let $\hat{z} = \boldsymbol{max}(z_1, \ldots., z_m)$: $\hat{z}$ is *sufficient* and the *mle* for θ *w.r.t.* the data $\boldsymbol{z}$.

Let us try to use the *EM* algorithm to compute the *mle* for θ given the *incomplete* (observed) data $(\boldsymbol{x},\boldsymbol{y})$, using the *complete* data $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$.

Now, for applying the EM algorithm we recall that:

$$log\ \mathbf{L}(\theta;\ t,\boldsymbol{x},\boldsymbol{y}) = log\ \mathbf{L}(\theta;\ t,\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) - log\ \boldsymbol{h}(\boldsymbol{z}\mid t,\boldsymbol{x},\boldsymbol{y},\theta).$$

But as t is ancillary and as $\boldsymbol{x}$ is function of $\boldsymbol{z}$ and t;

$\boldsymbol{z}$ is sufficient for θ *w.r.t.* data $(\boldsymbol{z},\boldsymbol{x},t)$,

so $$\mathbf{L}(\theta;\ t,\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \mathbf{L}(\theta;\ \boldsymbol{y},\boldsymbol{z}).$$

Evidently, the *mle* and the *sufficient statistic* for the complete data is:

$$\boldsymbol{argmax}_{\Theta}\ \boldsymbol{p}(t,\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\mid\theta) = \mathbf{max}\,(\hat{y},\hat{z}) = \hat{\theta}^*$$

as $\boldsymbol{p}(\boldsymbol{y},\boldsymbol{z}\mid\hat{\theta}^*,\theta) = [1/\hat{\theta}^*]^{n+m}$ for all $\theta \geq \hat{\theta}^*$

$= 0$ for all $\theta < \hat{\theta}^*$

independent of θ, for all θ consistent with the data, as properly summarized by the sufficient statistic $\hat{\theta}^*$ for the data.

For the *E-step* in *EM*

$$\boldsymbol{Q}(\theta \mid t,\boldsymbol{x},\boldsymbol{y},\hat{\theta}_j) = \int_{\boldsymbol{Z}} [\log \mathbf{L}(\theta; \boldsymbol{y},\boldsymbol{z})]\, \boldsymbol{h}(\boldsymbol{z} \mid t,\boldsymbol{x},\boldsymbol{y},\hat{\theta}_j)\, d\boldsymbol{z}$$

$$= \mathbf{E}_{t,\boldsymbol{x},\boldsymbol{y},\hat{\theta}_j} [\log \mathbf{L}(\theta; \boldsymbol{y},\boldsymbol{z})]$$

$$= \mathbf{E}_{t,\boldsymbol{x},\boldsymbol{y},\hat{\theta}_j} [\log [1/\theta]^{n+m}] \text{ for } \theta \geq \hat{\theta}^*$$

where $\hat{\theta}^* = \mathbf{max}\,(\hat{y},\hat{z})$,

which depends upon $\boldsymbol{x}$ only through $\hat{z}$ and upon $\boldsymbol{y}$ only through $\hat{y}$.

That is, $\log \mathbf{L}(\theta; \boldsymbol{y},\boldsymbol{z})] = \log [1/\theta]^{n+m}$

is constant in $(\boldsymbol{x},\boldsymbol{y})$ for each $\theta \geq \hat{\theta}^*$

So, for the *E*-step it appears that we require only to know

$$\mathbf{E}_{t,\boldsymbol{x},\boldsymbol{y},\hat{\theta}_j} [\hat{\theta}^*]$$

Observe that, as the z_i are conditionally *iid* given θ, and as x_i is a function only of z_i and the ancillary quantity t,

$$
\begin{aligned}
\boldsymbol{E}(z_i \mid t, \boldsymbol{x}, \boldsymbol{y}, \hat{\theta}_j) &= \boldsymbol{E}(z_i \mid t, \boldsymbol{x}, \hat{\theta}_j) \\
&= \boldsymbol{E}(z_i \mid t, x_i, \hat{\theta}_j) \\
&= \begin{cases} (1/2)(t+\hat{\theta}_j) & \text{if } x_i=1 \text{ (still running at time } t\text{)} \\ (1/2)t & \text{if } x_i=0 \text{ (not running at time } t\text{)} \end{cases}
\end{aligned}
$$

Thus, $$\mathbf{E}_{t,\boldsymbol{x},\boldsymbol{y},\hat{\theta}_j}[\hat{\theta}^*] = \mathbf{max}[\hat{y}, (1/2)(t+\hat{\theta}_j)],$$

as we have assumed that at least one $x_i = 1$, i.e., at least one of the m-disks is still spinning when we look at time t.

For the *M-step in EM* then we get:

$$\hat{\theta}_{j+1} = \boldsymbol{argmax}_{\Theta}\ \boldsymbol{Q}(\theta \mid t, \boldsymbol{x}, \mathbf{y}, \hat{\theta}_j)$$

$$= \boldsymbol{max}[\hat{y},\ (1/2)(t+\hat{\theta}_j)]$$

Thus, the *EM* algorithm iterates:

$$\hat{\theta}_{j+1} = \boldsymbol{max}[\hat{y},\ (1/2)(t+\hat{\theta}_j)]$$

and for each choice of $\hat{\theta}_0 > 0$,

$$lim_{j\to\infty}\ \hat{\theta}_{j+1} = \boldsymbol{max}[\hat{y}, t].$$

That is, the *EM* algorithm takes t to be sufficient for x, given that at least one of the m-disks is still spinning when we look at time t.

EM behaves here just as if $\hat{z} = t$.

Let $1 \leq k \leq m$ be the number of disks still spinning at time t, i.e. $k = \Sigma_i x_i$.

A more careful analysis of the likelihood function $\mathbf{L}(\theta;\, t,\boldsymbol{x},\boldsymbol{y})$ reveals that:

$$\begin{aligned}\mathbf{L}(\theta;\, t,\boldsymbol{x},\boldsymbol{y}) &= \mathrm{p}(\boldsymbol{y},\boldsymbol{x} \mid t,\theta) \\ &= \chi_{[\hat{y},\infty)}(\theta) \times \frac{1}{\theta}^{n} \times \frac{t}{\max(t,\theta)}^{m-k} \times \left(1 - \frac{t}{\max(t,\theta)}\right)^{k}\end{aligned}$$

So that:

$$\hat{\theta} = \boldsymbol{argmax}_{\Theta}\, \mathbf{L}(\theta;\, t,\boldsymbol{x},\boldsymbol{y}) \;=\; \boldsymbol{max}\left[\hat{y},\, \frac{n+m}{n+m-k}t\right]$$

and unless $\frac{n+m}{n+m-k}t \;\leq\; \hat{y}$,

$$\hat{\theta} \;>\; lim_{j\to\infty}\, \hat{\theta}_{j+1} = \boldsymbol{max}[\hat{y},t],$$

which is a larger value than the *EM* algorithm gives.

What goes wrong in the *EM* algorithm is that in computing the *E*-step, we have not attended to the important fact that the *log* likelihood function does not exist when $z_i > \theta$.

When computing $\mathbf{E}_{t,\boldsymbol{x},\boldsymbol{y},\hat{\theta}_j}$ [*log* $\mathbf{L}(\theta; \boldsymbol{y},\boldsymbol{z})$] at the $j+^{st}$ *E*-step, say, we use the fact that, given $x_i = 1$ and $\theta = \hat{\theta}_j$, then z_i is Uniform $\boldsymbol{U}[t, \hat{\theta}_j]$, with a conditional expected value of $(t+\hat{\theta}_j)/2$. However, for each parameter value θ, $t < \theta < \hat{\theta}_j$ with with positive $\hat{\theta}_j$-probability,

$$P_{t,x_i\hat{\theta}_j}(z_i\text{: } p(z_i \mid t,x_i,\theta) = 0\ ) > 0$$

and the expected *log*-likelihood for the *E*-step fails to exist for such θ!

The lesson to be learned from this example is this:

Before using the EM-algorithm, make sure that the log-likelihood function exists, so that the E-step is properly defined.

Additional References

Flury, B. and Zoppe, A. (2000) "Exercises in EM," *Amer. Staistican* **54**, 207-209.

Hastie, T., Tibshirani, R, and Friedman, J. *The Elements of Statistical Learning*. New York: Spring-Verlag, 2001, sections 8.5-8.6.