VE414 Lecture 18

Jing Liu

UM-SJTU Joint Institute

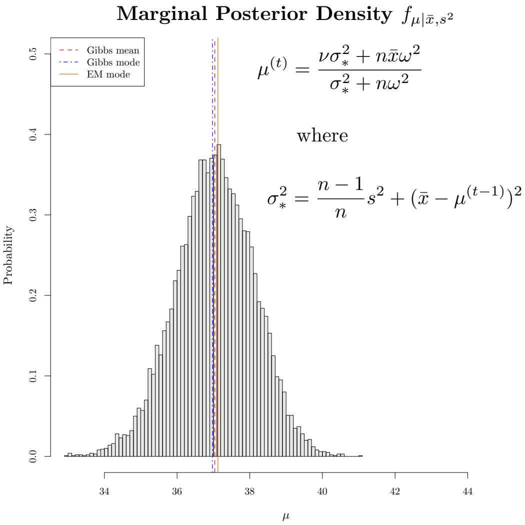
November 12, 2019

- The expectation maximisation (EM) is traditionally used for something else, here it can be used to find the mode of the marginal in a much simpler way.

```
> for (i in 1:9){
+
+   old = (xbar-data_mu[i])^2
+
+   sigma2_star = (n-1)/n * s2 + old
+
+   num = nu * sigma2_star + n * xbar * omega2
+   den = sigma2_star + n * omega2
+
+   data_mu[i+1] = num / den
+
+ }
> data_mu
```

```
 [1]  1.00000 22.64286 31.68842 35.75105 36.89255
 [6] 37.09146 37.12007 37.12404 37.12459 37.12466
```

- In just a few iterations, it produces a value similar to the one from Gibbs.



- In general, given a joint distribution up to a multiplicative constant A,

$$Af_{\mathbf{Y}|X} = Af_{\boldsymbol{\phi},\boldsymbol{\gamma}|X} = q_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}$$

where ϕ represents a subset of $\mathbf{Y}$ that we are interested in, i.e.

$$\mathbf{Y} = \begin{bmatrix}\boldsymbol{\phi} & \boldsymbol{\gamma}\end{bmatrix}^{\mathrm{T}}$$

obtaining the marginal is usually impossible, even up to a A^* is also difficult

$$A^* f_{\boldsymbol{\phi}|X} = q_{\boldsymbol{\phi}|X}$$

since it requires either finding the full conditional

$$f_{\boldsymbol{\phi}|X} \propto \frac{q_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}}; \qquad f_{\boldsymbol{\phi}|X} \not\propto \frac{q_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}}{q_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}}$$

or evaluating the following integral over the set $\mathcal{D}$ of all possible $\boldsymbol{\gamma}$

$$f_{\boldsymbol{\phi}} \propto \int_{\mathcal{D}} q_{\boldsymbol{\phi},\boldsymbol{\gamma}}(\boldsymbol{\phi},\boldsymbol{\gamma})\, d\boldsymbol{\gamma}$$

- Hence so far using a Monte Carlo method is our only viable option to obtain

$$\hat{\phi}$$

that is, a point estimate of $\phi \mid X$, mean, median or mode.

- The EM algorithm is a way to obtain the mode of the marginal without

$$f_{\boldsymbol{\phi}|X} \qquad \text{or} \qquad q_{\boldsymbol{\phi}|X}$$

in other words, it is an algorithm of maximising the marginal density without knowing the density function or the density function up to a constant!

- Consider the following identity, then logging the both sides, we have

$$f_{\boldsymbol{\phi}|X}(\boldsymbol{\phi} \mid x) = \frac{f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}(\boldsymbol{\phi},\boldsymbol{\gamma} \mid x)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)}$$

$$\ln\left(f_{\boldsymbol{\phi}|X}(\boldsymbol{\phi} \mid x)\right) = \ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}(\boldsymbol{\phi},\boldsymbol{\gamma} \mid x)\right) - \ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)\right)$$

- Taking the expectation on both sides, the term on the left reminds the same

$$
\begin{aligned}
\mathbb{E}\left[\ln\left(f_{\boldsymbol{\phi}|X}\left(\boldsymbol{\phi} \mid x\right)\right)\right] &= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\phi}|X}\left(\boldsymbol{\phi} \mid x\right)\right) f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x\right) d\boldsymbol{\gamma} \\
&= \ln\left(f_{\boldsymbol{\phi}|X}\left(\boldsymbol{\phi} \mid x\right)\right) \cdot 1
\end{aligned}
$$

and let the terms on the right become the following

$$
\begin{aligned}
\alpha\left(\boldsymbol{\phi}\right) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}\left(\boldsymbol{\phi},\boldsymbol{\gamma} \mid x\right)\right)\right] \\
&= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}\left(\boldsymbol{\phi},\boldsymbol{\gamma} \mid x\right)\right) f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x\right) d\boldsymbol{\gamma} \\
\beta\left(\boldsymbol{\phi}\right) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x\right)\right)\right] \\
&= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x\right)\right) f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x\right) d\boldsymbol{\gamma} \\
\ln\left(f_{\boldsymbol{\phi}|X}\left(\boldsymbol{\phi} \mid x\right)\right) &= \alpha\left(\boldsymbol{\phi}\right) - \beta\left(\boldsymbol{\phi}\right)
\end{aligned}
$$

the mode $\hat{\boldsymbol{\phi}}$ that maximises $f_{\boldsymbol{\phi}|X}$ if and only if $\hat{\boldsymbol{\phi}}$ maximises $\alpha\left(\boldsymbol{\phi}\right) - \beta\left(\boldsymbol{\phi}\right)$.

- Consider the following difference

$$\begin{aligned}\beta(\boldsymbol{\phi}) - \beta(\boldsymbol{\phi}^*) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)\right)\right] - \mathbb{E}\left[\ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)\right)\right] \\ &= \mathbb{E}\left[\ln\left(\frac{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)}\right)\right] \\ &= \int_{\mathcal{D}} \ln\left(\frac{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)}\right) f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)\, d\boldsymbol{\gamma}\end{aligned}$$

- Using Jensen's inequality, we have

$$\begin{aligned}\beta(\boldsymbol{\phi}) - \beta(\boldsymbol{\phi}^*) &\le \ln\left(\mathbb{E}\left[\frac{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)}\right]\right) \\ &= \ln\left(\int_{\mathcal{D}} f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)\, d\boldsymbol{\gamma}\right) = 0\end{aligned}$$

hence increasing/maximising $\alpha(\boldsymbol{\phi})$ increases/maximises $\alpha(\boldsymbol{\phi}) - \beta(\boldsymbol{\phi})$.

Algorithm 1: Expectation-Maximisation

Input : function $f_{\phi,\gamma|X}$, and $f_{\gamma|\{\phi,X\}}$, initial value $\phi^{(0)}$, tolerance ϵ
Output : mode ϕ_m

Function `EM(`$f_{\phi,\gamma|X}$, $f_{\gamma|\{\phi,X\}}$, $\phi^{(0)}$, ϵ`)`**:**
    $t \leftarrow 1$;
    while $t \leq 1e6$ **do**
        $\phi^{(t)} \leftarrow \arg\max_{\phi} \int_{\mathcal{D}} \ln\left(f_{\phi,\gamma|X}\left(\phi,\gamma \mid x\right)\right) f_{\gamma|\{\phi,X\}}\left(\gamma \mid \phi^{(t-1)}, x\right) d\gamma$
        if $\|\phi^{(t)} - \phi^{(t-1)}\| < \epsilon$ **then**
            $\phi_m \leftarrow \phi^{(t)}$;
            return ϕ_m ; `/* Solution */`
        else
            $t \leftarrow t + 1$;
        end if
    end while
    return *"Warning: 1 million iterations reached without achieving ϵ"* ;
end

- The EM algorithm essentially avoids one of the following two integrals

$$\int_{\mathcal{D}} q_{\boldsymbol{\phi},\boldsymbol{\gamma}}(\boldsymbol{\phi},\boldsymbol{\gamma})\, d\boldsymbol{\gamma} \qquad \text{or} \qquad \int_{\mathcal{D}} q_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma}\mid\boldsymbol{\phi},X)\, d\boldsymbol{\gamma}$$

  in return we are required to evaluate with the following integral

$$\begin{aligned}\alpha(\boldsymbol{\phi}) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x)\right)\right] \\ &= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x)\right) f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma}\mid\boldsymbol{\phi}^*,x)\, d\boldsymbol{\gamma}\end{aligned}$$

Q: Why is this a better deal in general? Because it looks a lot worse!

- Note $\boldsymbol{\phi}^{(t)}$ is the maximiser of α given a specific $\boldsymbol{\phi}^* = \boldsymbol{\phi}^{(t-1)}$ if and only if

$$\boldsymbol{\phi}^{(t)} = \arg\max_{\boldsymbol{\phi}} \int_{\mathcal{D}} \ln\left(q_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x)\right) q_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma}\mid\boldsymbol{\phi}^*,x)\, d\boldsymbol{\gamma}$$

- In addition to the above simplification, when the full conditional distribution $f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}$ is available, the EM often reduces to simple iterative evaluation.

- In terms of the following model,

$$\begin{aligned} X \mid \{\mu, \sigma^2\} &\sim \text{Normal}\left(\mu, \sigma^2\right) \\ \mu &\sim \text{Normal}\left(\nu, \omega^2\right) \\ \sigma^2 &\sim \varphi_{\sigma^2} \end{aligned}$$

we have derived the followings last time

$$q_{\mu,\sigma^2}\left(\mu, \sigma^2\right) = \left(\sigma^2\right)^{-(1+n/2)} \cdot \exp\left(-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2} - \frac{(\mu-\nu)^2}{2\omega^2}\right)$$

$$f_{\sigma^2|\{\mu,\bar{x},s^2\}} = \text{Scaled Inverse}\,\chi^2\left(n, \frac{(n-1)s^2}{n} + (\bar{x}-\mu)^2\right)$$

- Hence within each iteration, we have to maximise the following w.r.t μ

$$\begin{aligned} \alpha(\mu) &= \mathbb{E}\left[\ln\left(f_{\{\mu,\sigma^2\}|\{\bar{x},s^2\}}\left(\mu, \sigma^2 \mid \bar{x}, s^2\right)\right)\right] \\ &= \mathbb{E}\left[-(2+n)\ln\sigma - \frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] - \frac{(\mu-\nu)^2}{2\omega^2} - \ln A \end{aligned}$$

- Rearranging into the following form,

$$\alpha(\mu) = \mathbb{E}\left[-(2+n)\ln\sigma - \frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] - \frac{(\mu-\nu)^2}{2\omega^2} - \ln A$$

$$= -\frac{1}{2}\mathbb{E}\left[\frac{1}{\sigma^2}\right]\left((n-1)s^2 + n(\bar{x}-\mu)^2\right) - \frac{(\mu-\nu)^2}{2\omega^2}$$

$$\underbrace{-(2+n)\mathbb{E}\left[\ln\sigma\right] - \ln A}_{\text{additive constant w.r.t. } \mu}$$

$$= -\frac{1}{2}\mathbb{E}\left[\frac{1}{\sigma^2}\right]\left((n-1)s^2 + n(\bar{x}-\mu)^2\right) - \frac{(\mu-\nu)^2}{2\omega^2} + \text{constant}$$

- Recall the expectation is over σ^2 given $\mu^* = \mu^{(t-1)}$, $\bar{x}$ and s^2, which means

$$\sigma^2 \mid \{\mu^{(t-1)}, \bar{x}, s^2\} \sim \text{Scaled Inverse}\,\chi^2\left(n, \frac{(n-1)s^2}{n} + (\bar{x}-\mu^{(t-1)})^2\right)$$

$$\mathbb{E}\left[\frac{1}{\sigma^2}\right] = \left(\frac{(n-1)s^2}{n} + (\bar{x}-\mu^{(t-1)})^2\right)^{-1}$$

- Thus, in each iteration, we need to solve the following

$$\mu^{(t)} = \arg\max_{\mu} \left\{ \mathbb{E}\left[\ln\left(f_{\{\mu,\sigma^2\}|\{\bar{x},s^2\}} \left(\mu, \sigma^2 \mid \bar{x}, s^2 \right) \right) \right] \right\}$$

$$= \arg\max_{\mu} \left\{ -\frac{\left((n-1)s^2 + n(\bar{x}-\mu)^2 \right)}{2\sigma_*^2} - \frac{(\mu-\nu)^2}{2\omega^2} + \text{constant} \right\}$$

where $\sigma_*^2 = \dfrac{(n-1)s^2}{n} + (\bar{x} - \mu^{(t-1)})^2$.

Q: Have you seen this before?

$$q_\mu \propto \exp\left(-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2} - \frac{(\mu-\nu)^2}{2\omega^2} \right)$$

which is the unnormalised posterior of μ when σ^2 is known and normal prior $\mathrm{Normal}\left(\nu, \omega^2\right)$ is used, the posterior is know to be

$$\mu \mid \{\sigma^2, \bar{x}, s^2\} \sim \mathrm{Normal}\left(\frac{\omega^2\bar{x} + \nu\sigma^2/n}{\omega^2 + \sigma^2/n}, \frac{\omega^2\sigma^2/n}{\omega^2 + \sigma^2/n} \right)$$

- Therefore, the solution to the maximisation in each iteration is simply

$$\mu^{(t)} = \frac{n\omega^2\bar{x} + \nu\sigma_*^2}{n\omega^2 + \sigma_*^2} \qquad \text{where} \quad \sigma_*^2 = \frac{(n-1)s^2}{n} + (\bar{x} - \mu^{(t-1)})^2$$

since the objective function of the maximisation

$$-\frac{\left((n-1)s^2 + n(\bar{x}-\mu)^2\right)}{2\sigma_*^2} - \frac{(\mu-\nu)^2}{2\omega^2} + \text{constant}$$

corresponds to the logarithm of the normal density,

$$\mathrm{Normal}\left(\frac{\omega^2\bar{x} + \nu\sigma_*^2/n}{\omega^2 + \sigma_*^2/n}, \frac{\omega^2\sigma_*^2/n}{\omega^2 + \sigma_*^2/n}\right)$$

for which we know the maximum happens at where the mean is.

- Using this iterative formula recursively, we reach the the maximiser of

$$f_{\mu|\{\sigma^2,\bar{x},s^2\}}$$

- This leads to what I have used and shown you in the beginning.