Lecture 21 Math 390.03-02 5/19/16

1977 Dempster, Laird, Rubin

But 1974 Rolf Sundberg found it first

1977 - generalized method

Wu, 1983: proved convergence for a wide variety of parametric models

---

Newton Raphson: useful for solving $f(x) = c$. We used it for finding $\hat{\theta}_{MAP} = \text{argmax}\{k(\theta|x)\}$ by setting $k'(\theta|x) = 0$ when it was not soluble in closed form.

E-M: useful for cases where MLE's break down due to not in closed form but if you knew some latent data it would be easy.

Can this help with our semi-conjugate model?

$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$, $\theta \sim N(\mu_0, \tau^2)$, $\sigma^2 \sim \text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0 \sigma_0^2}{2}\right)$

$$P(\theta, \sigma^2 | X) \propto P(X|\theta, \sigma^2)\, P(\theta)\, P(\sigma^2)$$

$$\propto \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i - \theta)^2}\right) \left(e^{-\frac{1}{2\tau^2}(\theta - \mu_0)^2}\right) \left((\sigma^2)^{-\frac{n_0}{2} - 1} e^{-\frac{n_0 \sigma_0^2}{2\sigma^2}}\right)$$

$$\propto (\sigma^2)^{-\frac{n_0 + n}{2} - 1} e^{-\frac{\sum x_i^2 + n_0 \sigma_0^2}{2\sigma^2}} \underbrace{e^{\frac{\theta n \bar{x}}{\sigma^2}} e^{-\frac{n\theta^2}{2\sigma^2}} e^{-\frac{\theta^2}{2\tau^2}} e^{\frac{\theta \mu_0}{\tau^2}}}$$

$$e^{\underbrace{\left(\frac{\mu_0}{\tau^2}+\frac{n\bar{x}}{\sigma^2}\right)}_{a}\theta + \underbrace{\left(-\frac{n}{2\sigma^2}-\frac{1}{2\tau^2}\right)}_{b}\theta^2}$$

Need $e^{-\frac{1}{2v}(\theta-c)^2}$

$$\Rightarrow -\frac{1}{2v}(\theta-c)^2 = -\frac{1}{2v}(\theta^2-2\theta c+c^2) = -\frac{\theta^2}{2v}+\frac{\theta c}{v}-\frac{c^2}{2v}$$

$$\Rightarrow -\frac{1}{2v} = b \Rightarrow v = -\frac{1}{2b} = -\frac{1}{2\left(-\frac{n}{2\sigma^2}-\frac{1}{2\tau^2}\right)} = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$$

$$\frac{c}{v} = a \Rightarrow c = av = \frac{\frac{\mu_0}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$$

$$-\frac{c^2}{2v} = bc^2 = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\left(\frac{\frac{\mu_0}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2 = Q \quad \text{s.t. } Q \text{ is a } f(\sigma^2)$$

$$\sqrt{2\pi v}\underbrace{\frac{1}{\sqrt{2\pi v}}e^{a\theta}e^{b\theta^2}e^{Q}}e^{-Q}$$

$$= N(\theta_p, \sigma_p^2)$$

$$\Rightarrow P(\theta,\sigma^2|x) \propto N(\theta_p,\sigma_p^2)\underbrace{\overbrace{(\sigma^2)^{-\frac{n_0+n}{2}-1}e^{-\frac{\Sigma x_i^2+n_0\sigma_0^2}{2\sigma^2}}}^{\text{kernel of inv gamma}}\overbrace{e^{-Q}\sqrt{\frac{2\pi}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}}}^{\text{other stuff}}}_{K(\sigma^2|x)}$$

Resort to grid sampling of $K(\sigma^2|x)$ to approximate $P(\sigma^2|x)$ and then sample σ^2, sample $N(\theta_p,\sigma_p^2)$

Can we do better? Is there an [illegible] algorithm?

$P(\theta, \sigma^2 | X)$ non-std distr.

But $P(\theta | X, \sigma^2) = N(\theta_p, \sigma^2_p)$

$\Sigma(x_i - \theta)^2 = n\hat{\sigma}^2$

$$P(\sigma^2 | X, \theta) \propto P(X | \theta, \sigma^2)\, P(\sigma^2)$$

$$\propto (\sigma^2)^{-\frac{n_0+n}{2}-1}\, e^{-\frac{\Sigma x_i^2 - n_0\sigma_0^2 + n\theta^2 - 2\theta n \bar{x}}{2\sigma^2}}$$

$$\propto \text{InvGamma}\left(\frac{n_0+n}{2}, \frac{n_0\sigma_0^2 + n\hat{\sigma}^2}{2}\right)$$

But in $\theta | X, \sigma^2 \Rightarrow \sigma^2$ is unknown and

in $\sigma^2 | X, \theta \Rightarrow \theta$ is unknown!

Algorithm: similar to E-M:

Step 1: Guess σ_0^2 ... maybe use s^2

Step 2: Sample $\theta | X, \sigma^2 = \sigma_0^2$ to get θ_0

Step 3: Sample $\sigma^2 | X, \theta = \theta_0$ to get σ_1^2

Step 4: Sample $\theta | X, \sigma^2 = \sigma_1^2$ to get θ_1

Step 5: Repeat Steps 3 & 4 to obtain...

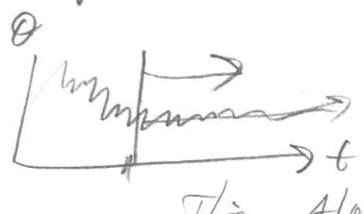
$$\left\langle \begin{bmatrix}\theta_0 \\ \sigma_0^2\end{bmatrix}, \begin{bmatrix}\theta_1 \\ \sigma_1^2\end{bmatrix}, \ldots, \begin{bmatrix}\theta_t \\ \sigma_t^2\end{bmatrix} \right\rangle$$

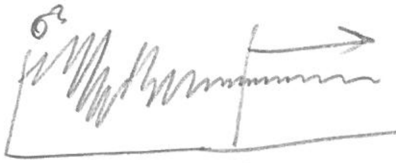
where $\langle \theta_0, \theta_1, \ldots \rangle$ and $\langle \sigma_0^2, \sigma_1^2, \ldots \rangle$ form two "chains"

where t is large

Quit upon convergence of the chains max t s.t. all converge

Traceplots:

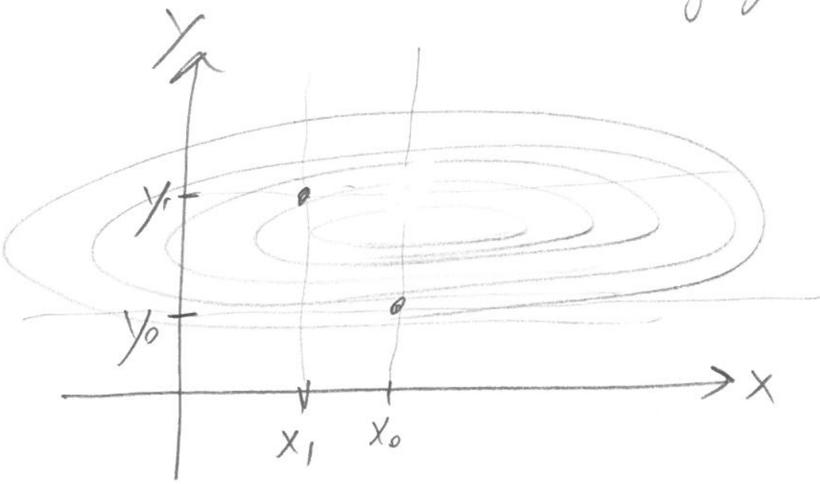




$B = \max_t \{\text{both converge}\}$

This Algorithm is called "Gibbs Sampling"

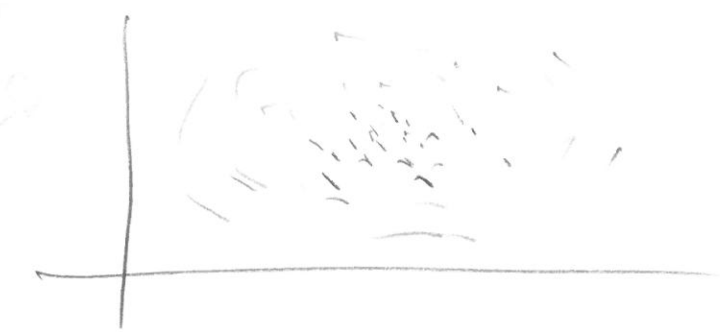
What is going on?



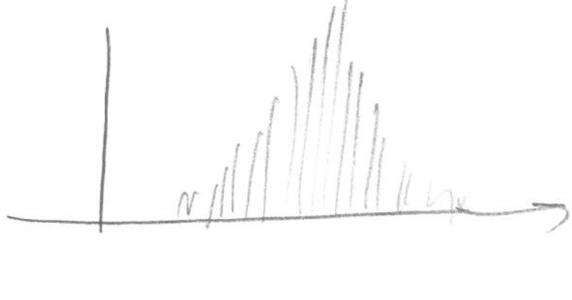
$f(X,Y)$ is a density you wish to sample from but cannot. But... You know $f(x|y)$ & $f(y|x)$

So begin w/ y_0: You sample $x_0 = f(x|y=y_0)$. Then you sample $y_1 = f(y|X=x_0)$ then $x_1 = f(x|y=y_1)$, etc.

Eventually...



You sample the whole space. If you only care about x, then you ditch the y's and have an estimate of $f(x)$:

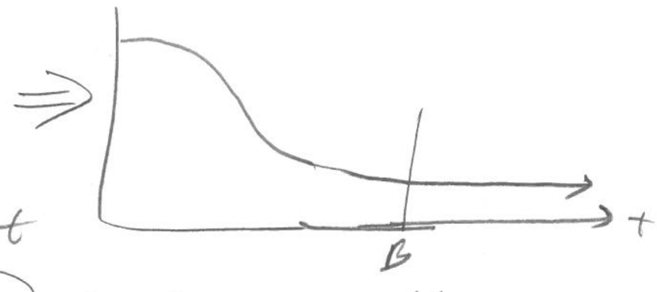


Problem #1 At what pt. did it converge?

Trace plot

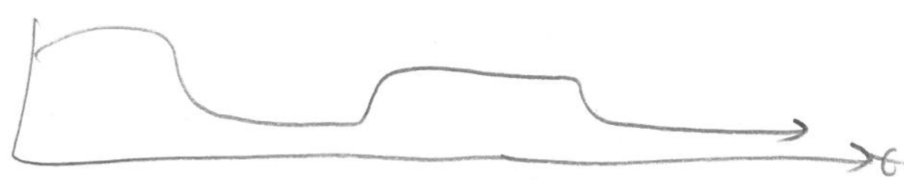
θ_t

running avg of θ_t



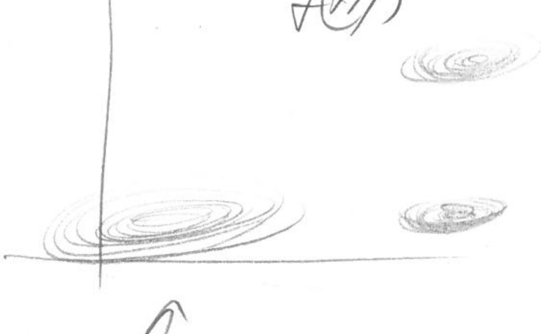
It appears to converge @ $t \geq B$. "Good mixing"

But did it really converge?



If you let it run long enough you may see this, corresponding to bad "mixing". Mixing is the ability of the chain to efficiently traverse the param space (i.e. the support).

$f(x,y)$



Hard density to Gibbs sample

Problem #2 If $f(x,y)$ has multiple modes, the Gibbs chain can get stuck in a mode and never get out.

Diagnosis.. run lots of chains from multiple starting positions and look at traceplots. As $\dim(\theta)$ increases... this becomes a bigger problem