Lecture 15 Math 390.03-01 4/4/16

L1

Semi Conjugacy

Situation

You know there's no dependency

$\sigma^2 \perp \theta$ a priori $P(\theta, \sigma^2) = P(\theta) P(\sigma^2)$

$\theta \sim N(150, 10^2)$, σ^2 unknown so... $\sigma^2 \sim$ Jeffreys

when $P(\sigma^2)$ is conj prior for $P(\sigma^2 | X, \theta)$

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

but $P(\theta, \sigma^2)$ is <u>not</u> conj prior for $P(\theta, \sigma^2 | X)$

$\theta \sim N(\mu_0, \tau^2)$ $\quad \sigma^2 \sim \text{InvGamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$

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

$$\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{X}-\theta)^2\right)} e^{-\frac{1}{2\tau^2}(\theta - \mu_0)^2} (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} e^{-\frac{\nu_0 \sigma_0^2}{2} / \sigma^2}$$

$$\propto e^{-\frac{n}{2\sigma^2}(\bar{X}-\theta)^2} e^{-\frac{\theta^2}{2\tau^2}} e^{\frac{\theta \mu_0}{\tau^2}} (\sigma^2)^{-\left(\frac{\nu_0 + n}{2}+1\right)} e^{-\frac{\nu_0 \sigma_0^2 + (n-1)s^2}{2} / \sigma^2}$$

$$= e^{-\frac{n}{2\sigma^2}\bar{X}^2} e^{\frac{n\bar{X}\theta}{\sigma^2}} e^{-\frac{n\theta^2}{2\sigma^2}} \quad \text{"} \quad \text{"} \quad \text{"}$$

$$= e^{\left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)\theta - \left(\frac{1}{2\tau^2} + \frac{n}{2\sigma^2}\right)\theta^2} (\sigma^2)^{-\left(\frac{\nu_0+n}{2}+1\right)} e^{-\frac{\nu_0 \sigma_0^2 + (n-1)s^2 + n\bar{X}^2}{2} / \sigma^2}$$

$\uparrow$ b $\quad \uparrow$ a

$-\frac{1}{2v} = a$, $\quad d = b$

$\Rightarrow v = -\frac{1}{2a} = \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$

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

$= k(\sigma^2 | X) \propto P(\sigma^2 | X)$ ~ non-std.

$$\propto e^{-\frac{1}{2v}(\theta - d)^2} e^{+\frac{d^2}{2v}} \quad \text{"} \quad \text{"} \quad \text{"}$$

$\Rightarrow$ not an inv. gamma

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

$$\propto N(\theta_p, \sigma_p^2)\, e^{\frac{1}{2}\frac{\left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)^2}{\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)}} = \frac{\frac{(\tau^2 n\bar{X} + \sigma^2\mu_0)^2}{(\sigma^2\tau^2)^2}}{\frac{\sigma^2 + n\tau^2}{\sigma^2\tau^2}} \quad K e^{\frac{1}{\sigma^2}}$$

$\Rightarrow$ what do you do?

$$P(\theta, \sigma^2 | x) \propto \overset{\text{HW}}{\downarrow}{N(\theta_p, \sigma_p^2)}\; \underbrace{k(\sigma^2 | x)}_{\text{not a distribution ... can't sample from it}}$$

How about the following idea... Finite n approx. No $n \to \infty$ req'd!

$$P(\theta | x) = c(x)\, k(\theta | x)$$

$$\ln P(\theta | x) = \ln c(x) + \underbrace{\ln(k(\theta | x))}_{g(\theta | x)}$$

Recall $f(x) \approx \text{Tay}(\underset{\text{free var}}{x}, \underset{\text{center}}{c}, \underset{\text{degree}}{d}) := \sum_{i=0}^{d} \frac{f^{(i)}(c)}{i!}(x-c)^i$

$$g(\theta | x) \approx \text{Tay}(\theta, c, 2) = g(c|x) + g'(c|x)(\theta - c) + \frac{g''(c|x)(\theta - c)^2}{2}$$

What should c be?

Recall $\hat{\theta}_{MAP} := \text{argmax}\{P(\theta|x)\} = \text{argmax}\{\ln P(\theta | x)\} = \text{argmax}\{\underbrace{\ln k(\theta|x)}_{g(\theta|x)}\} =$

And $g'(\hat{\theta}_{MAP} | x) = 0$

let $c = \hat{\theta}_{MAP}$

$$\Rightarrow g(\theta | x) \approx g(\hat{\theta}_{MAP} | x) + \underbrace{\cancel{g'(\hat{\theta}_{MAP} | x)(\theta - \hat{\theta}_{MAP})}}_{\to 0} + \frac{1}{2} g''(\hat{\theta}_{MAP} | x)(\theta - \hat{\theta}_{MAP})^2$$

$$\Rightarrow \ln P(\theta | x) \approx \ln(c(x)) + g(\hat{\theta}_{MAP} | x) + \frac{1}{2} g''(\hat{\theta}_{MAP} | x)(\theta - \hat{\theta}_{MAP})^2$$

$$\Rightarrow P(\theta | x) \approx c(x)\, e^{g(\hat{\theta}_{MAP} | x)}\, e^{\frac{1}{2} g''(\hat{\theta}_{MAP}|x)(\theta - \hat{\theta}_{MAP})^2}$$

$$\propto \quad \propto N\left(\hat{\theta}_{MAP}, \left(\frac{1}{\sqrt{g''(\hat{\theta}_{MAP}|x)}}\right)^2\right)$$

Now you can sample from $k(\theta | x)$!

Does this make sense?

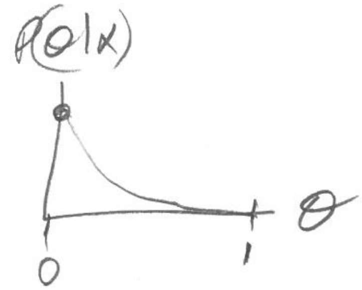
$E[\theta|x] \approx \hat{\theta}_{MAP}$ yes...

Recall

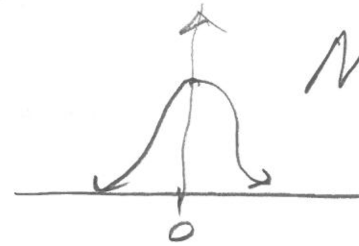
$\theta \sim Beta(1,1)$

$x = 0,\ n = 3$

$\theta|x \sim Beta(1, 4)$



$\overset{?}{\approx}$



NO...

Unless you really think the shape of the posterior is normal... all bets are off!!

Next idea

$P(\theta, \sigma^2|x) \propto N(\theta_p, \sigma_p^2)\, k(\sigma^2|x)$

Let's do the following... create a grid $\{\theta_1, \ldots, \theta_G\}$ evenly spaced

1) And calculate $\hat{c} := \left(\sum_{g=1}^{G} k(\theta_g|x)\right)^{-1}$

then use...

$P(\theta_g|x) \approx \hat{c}\, k(\theta_g|x)$

In practice... keep all $k(\theta_g|x)$ to create $\hat{F}(\theta_i|x) = \sum_{g=1}^{i} \hat{c}\, k(\theta_g|x)$

Draw $U \sim U(0,1)$ and return $\hat{F}^{-1}(U|x)$

Disadvantages

① Numerically unstable.. $k(\theta|x) = 0$ or ∞ in a computer

Sol: Sample $\ln k(\theta|x)$ and exponentiate afterwards

② If $Supp(\theta)$ is an unbounded set, what support subset should we use?

$$\sigma_g^2 \in [0.0001, 1{,}000{,}000] \,?$$

need to make these decisions

If you make a wrong decision... you may miss some of the effective support.

(In our case, we know $\sigma^2 | x$ is unimodal, so we can end the grid when $k(\sigma_g^2|x)$ gets small as $\sigma_g^2 \uparrow$ and we can start close to 0.)

③ In multiple dimensions, unfeasible. Even 100^{10} is ∞ in a computer.

For now: grid sampling okay... but we will teach new ideas soon! We will take a break from Bayesian stuff to discuss the linear model: bedrock of all stat. modeling.

Consider draws from a bivariate distribution X, Y

$\langle x_1, y_1 \rangle$
$\langle x_2, y_2 \rangle$
$\vdots$
$\langle x_n, y_n \rangle$

WLOG, let y be the
"response", "outcome" or "dependent" variable
and x is the
"feature", "covariate", "regressor", "independent" variable

$$X \xrightarrow[\text{change}]{\text{affects a}} Y$$

which may or may not be a causal effect

Causality beyond scope of course

$$X \xrightarrow{f, \varepsilon} Y$$

there is some function f and some noise generation ε (epsilon)

$$y = f(x) + \varepsilon$$

The goal is to model f.

$f(x) \in \mathcal{F}$ there's a big space out there!

But for our purposes now... restrict

$$f(x) \in \mathcal{F}_{lin} := \{ \beta_0 + \beta_1 x : \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R} \}$$

And restrict

$\varepsilon \neq h(x)$, noise is independent

Before we get into random variables...