Lecture 15 Math 390.03-4 4/4/16

Semi Conjugacy

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

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

Situation

You know there's no dependence

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

$\theta \sim N(\mu_0, \tau^2) \qquad \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$

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

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

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

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

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

(b: coefficient of θ; a: coefficient of θ^2)

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

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

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

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

$$\Rightarrow P(\theta | x, \sigma^2) \propto N\left(\theta_p, \sigma_p^2\right) e^{\frac{1}{2}\frac{\left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)^2}{\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)}} = \frac{\left(\tau^2 n\bar{x} + \sigma^2\mu_0\right)^2}{\frac{(\sigma^2\tau^2)^2}{\sigma^2 + n\tau^2}\,\sigma^2\tau^2} \propto e^{\frac{1}{\sigma^2}}$$

$$= K(\sigma^2 | x) \propto P(\sigma^2 | x) \text{ non-std.}$$

$\Rightarrow$ not my Inv. gamma

$\Rightarrow$ What do you do?

this derivation is mostly right... but may be off by a bit

HW

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

has a distribution ... can 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 linear combo of derivs

$$f(x) \approx \text{Tay}(x, c, d) := \sum_{i=0}^{d} \frac{f^{(i)}(c)}{i!} (x-c)^i$$

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

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

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

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

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

$$\propto \leftarrow \propto N\left(\hat{\theta}_{map}, \left(\sqrt{-g''(\hat{\theta}_{map}|x)}\right)^{-2}\right)$$

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

Does this make sense?

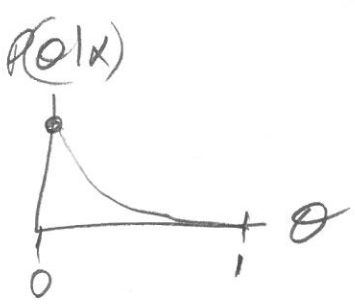
$E[\theta|x] \approx \hat{\theta}_{MAP}$ yes...

Recall

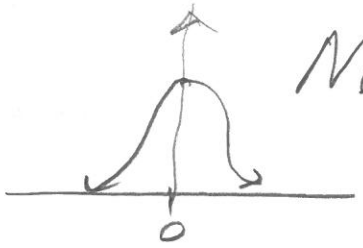
$\theta \sim Beta(1,1)$

$x = 0, n = 3$

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



$\approx$?



Unless your 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 $\text{Supp}(\theta)$ 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 learn two 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

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

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

$X \xrightarrow[\text{change}]{\text{affects}} Y$ Changes 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 generator 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...