Lecture 15

Let $X_1, \ldots, X_n \overset{\text{exch}}{\sim} N(\theta, \sigma^2)$

We were looking at the posterior $P(\theta, \sigma^2 | x)$.

We had

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

under the informative prior, $p(\theta, \sigma^2) \propto N\left(\mu_0, \frac{\sigma^2}{m}\right) \text{InvGamma}\left(\frac{v_0}{2}, \frac{v_0 \sigma_0^2}{2}\right)$

and our posterior was

$$p(\theta, \sigma^2 | x) \propto N\left(\frac{n}{n+m}\bar{x} + \frac{m}{n+m}\mu_0, \frac{\sigma^2}{n+m}\right) \text{InvGamma}\left(\frac{v_0+n}{2}, \frac{v_0\sigma_0^2 + (n-1)s^2 + \frac{nm}{n+m}(\bar{x}-\mu_0)^2}{2}\right)$$

Thus we have

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

$$\text{NIG} \propto \text{Normal Likelihood} \cdot \text{NIG}$$

★ So the NIG is the conjugate prior for the normal likelihood ★

Now in the above case, $p(\theta, \sigma^2) = \underbrace{p(\theta | \sigma^2)}_{N\left(\mu_0, \frac{\sigma^2}{m}\right)}\, \underbrace{p(\sigma^2)}_{\text{InvGamma}\left(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2}\right)}$

so the distribution of θ depends on σ^2.

What happens if $p(\theta, \sigma^2) = p(\theta)\, p(\sigma^2)$. I.e. if θ & σ^2 are independent?

Let $p(\theta, \sigma^2) = \underbrace{p(\theta)}\, p(\sigma^2)$

$$= N(\mu_0, \tau^2)\ \text{InvGamma}\left(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2}\right)$$

where $\tau^2 \neq f(\sigma^2)$.

Then in this case we have

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

$$= p(x | \theta, \sigma^2)\, p(\theta)\, p(\sigma^2)$$

$$= \text{NORMAL LIK} \cdot N(\mu_0, \tau^2) \cdot \text{InvGamma}\left(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2}\right)$$

$$\propto \underbrace{N(\theta_p, \sigma_p^2)} \cdot \text{stuff}$$

$$\propto p(\theta | x, \sigma^2) \cdot k(\sigma^2 | x)$$

BUT $k(\sigma^2 | x) \not\propto$ InvGamma &

$k(\sigma^2 | x) \not\propto$ any known distribution

Now $p(\theta, \sigma^2 | x) \propto N(\theta_p, \sigma_p^2) \cdot$ UNKNOWN DISTRIBUTION.

This leads to problems! How can we sample from it? It's not a distribution we know so we can't sample from it. We don't know what our posterior is!

Now pretend you only have kernel. So you only have $k(\sigma^2|x)$. By definition of kernel,

$$p(\sigma^2|x) = c(x) \cdot k(\sigma^2|x)$$

Remember, we only have $k(\sigma^2|x)$. If we can estimate $c(x)$ then we can estimate $p(\sigma^2|x)$ & hence estimate our posterior $p(\theta, \sigma^2|x)$.

$$p(\sigma^2|x) = c(x) \cdot k(\sigma^2|x)$$

$$\Rightarrow \ln p(\sigma^2|x) = \ln c(x) + \ln k(\sigma^2|x)$$

Call $g(\sigma^2|x) := \ln k(\sigma^2|x)$ so

$$\ln p(\sigma^2|x) = \ln c(x) + g(\sigma^2|x)$$

Now,

$$g(\sigma^2|x) \approx \text{Tay}(\sigma^2, c, 2) = \text{Tay}(\sigma^2, c, 2)$$

$$= \sum_{i=0}^{2} \frac{g^i(c)}{i!} \cdot (x-c)^i$$

$$= g(c|x) + g'(c|x)(\sigma^2 - c) + \frac{g''(c|x)(\sigma^2-c)^2}{2}$$

Recall a function $f(x)$ can be approximated by a Taylor series denoted $\text{Tay}(x, c, d)$ (variable, center, degree)

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

So we get a function which does NOT RELY on σ^2. Instead this relies on c. For what value should we choose c to be? Let $c = \hat{\sigma}^2_{MAP}$. Now

$\hat{\theta}_{MAP} := \text{argmax}\{p(\sigma^2|x)\} = \text{argmax}\{k(\sigma^2|x)\} = \text{argmax}\{\ln(k(\sigma^2|x))\}$

So $g'(\hat{\sigma}^2_{MAP}|x) = 0$ b/c $\hat{\sigma}^2_{MAP}$ is the maximum. Hence for $c = \hat{\sigma}^2_{MAP}$,

$$g(\hat{\sigma}^2_{MAP}|x) + g'(\hat{\sigma}^2_{MAP}|x)(\sigma^2 - \hat{\sigma}^2_{MAP}) + \frac{g''(\hat{\sigma}^2_{MAP}|x)(\sigma^2 - \hat{\sigma}^2_{MAP})^2}{2}$$

$$= g(\hat{\sigma}^2_{MAP}|x) + g''(\hat{\sigma}^2_{MAP}|x)(\sigma^2 - \hat{\sigma}^2)^2 \cdot \tfrac{1}{2}$$

$$= \text{Tay}(\sigma^2, \hat{\sigma}^2_{MAP}, 2) \approx g(\sigma^2|x)$$

Now $g(\sigma^2|x) \approx g(\hat{\sigma}^2_{MAP}|x) + g''(\hat{\sigma}^2_{MAP}|x)(\sigma^2 - \hat{\sigma}^2)^2 \cdot \frac{1}{2}$. Now,

$$p(\sigma^2|x) \propto k(\sigma^2|x) = e^{g(\sigma^2|x)} \approx e^{g(\hat{\sigma}^2_{MAP}|x)} \; e^{\frac{1}{2g''(\hat{\sigma}^2_{MAP}|x)^{-1}}(\sigma^2 - \hat{\sigma}^2_{MAP})^2}$$

$$\propto e^{\frac{1}{2g''(\hat{\theta}_{MAP}|x)^{-1}}(\sigma^2 - \hat{\sigma}^2_{MAP})}$$

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

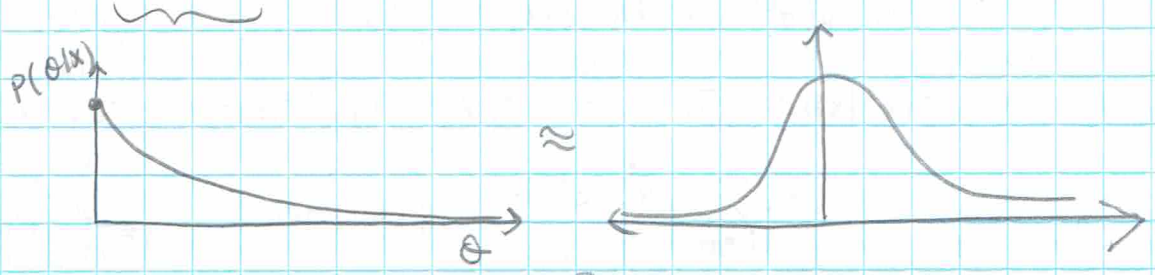
So we have the approximate distribution of $\sigma^2|x$! YAYYYY

More generally, $k(z|x) \propto N\left(\hat{z}_{MAP}, \left(\frac{1}{\sqrt{g''(\hat{z}_{MAP}|x)}}\right)^2\right)$

where $g(z|x) = \ln k(z|x)$.

This approximation can be bad! Recall if $X \sim Bin(n,\theta)$, $\theta \sim Beta(1,1)$ then $\theta|x \sim Beta(x+\alpha, (n-x)+\beta)$. For $x=0, n=3$, we get $\theta|x \sim Beta(1,4)$. Note $\hat{\theta}_{MAP} = 0$

So $\theta|x \sim Beta(1,4)$ & by above, $\theta|x \approx N(0, \text{something})$ so the Beta(1,4) distribution $\approx N(0, \text{something})$



which is "false"

What I really mean is that the error is huge!!!

Now remember $P(\theta, \sigma^2|x) \propto \underbrace{p(\theta|x, \sigma^2)}_{\text{KNOWN}} \underbrace{k(\sigma^2|x)}_{\text{KNOWN what the function is but NOT know what the distribution is. (i.e. you don't know what the p.d.f. is)}}$

We want to draw/sample from $p(\theta, \sigma^2|x)$ but we need to know the distribution of $k(\sigma^2|x)$.

Here is the GRID METHOD.

① Let $D \subseteq$ support of σ^2

② Fix $\varepsilon > 0$

③ You know the function $k(\sigma^2|x)$

④ Define
$$\sigma_1^2 = \min(D)$$
$$\sigma_2^2 = \min(D) + \varepsilon$$
$$\sigma_3^2 = \min(D) + 2\varepsilon$$
$$\cdots$$
$$\sigma_n^2 = \max(D)$$

⑤ You get $\{\sigma_1^2, \ldots, \sigma_n^2\}$

⑥ Plugging into k yields $\{k(\sigma_1|x), \ldots, k(\sigma_n|x)\}$

⑥ $P(\sigma_j^2|x) \approx \hat{c}\, k(\sigma_j|x)$ where $\hat{c} = \frac{1}{\sum_{j=1}^{n} k(\sigma_j|x)}$

⑦ So $\hat{F}(\sigma_j^2|x) = \sum_{j=1}^{k} \hat{c}\, k(\sigma_j^2|x)$

⑧ $\sigma^2 = \hat{F}^{-1}(u)$

Problems with GRID APPROX

1. If ε is too big
2. If D is too small
3. If the kernel is multidimensional, then problems in estimating kernel.

Now lets discuss a time before probability, r.v.'s etc.

Consider a bivariate distribution X, Y

Let Y be the "outcome", "response," "dependent variable"

Let X be the "regressor," "feature," "covariate," "independent variable"

$$X \xrightarrow{\text{affects}} Y$$

How does X affect Y? By some function & by some noise

Draws from a bivariate distribution look like

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

which is a 2-dimensional sample.

$y = f(x) + \varepsilon$ where $\varepsilon \neq h(x)$

primary goal: obtain $f(x)$

Secondary goal: obtain ε

Let $\mathcal{F} = \{ f \mid f: \mathbb{R} \to \mathbb{R} \}$

This set is too large, so limit the size of $\mathcal{F}$.

$$\mathcal{F}_0 = \{ \beta_0 + \beta_1 x : \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R} \}$$