JAIVARDHAN KAPOOR
ROLL NO: 150300

CS698S - Homework 2

PROBLEM 1

given $\vec{y} = (y_1, y_2 \ldots\ldots y_N) \sim p(y_n | \theta)$

and $\hat{\theta}$ is a point estimate of θ, with posterior $p(\theta | \vec{y})$ over θ.

with point estimate,

probability of $y_* = \mu_1 = p(y_* | \hat{\theta})$

with complete posterior,

predictive posterior of $y_* = \mu_2 = \int p(y_* | \theta) p(\theta | \vec{y}) d\theta = p(y_* | \vec{y})$

now, $E_{p(y_*|\vec{y})}[l(y_*, \mu_1)] - E_{p(y_*|\vec{y})}[l(y_*, \mu_2)]$

$$= E_{p(y_*|\vec{y})}[l(y_*, \mu_1) - l(y_*, \mu_2)]$$

$$= \int (-\log \mu_1 + \log \mu_2)\, p(y_* | \vec{y})\, dy^*$$

$$= \int -\log \frac{\mu_1}{\mu_2} p(y_* | \vec{y}) dy^* = \int -\log \left[\frac{p(y_* | \hat{\theta})}{p(y_* | \vec{y})}\right] p(y_* | \vec{y}) dy^*$$

$$= KL\left(p(y_* | \vec{y}) \,||\, p(y_* | \hat{\theta})\right) \geq 0$$

Therefore, $E_{p(y_*|\vec{y})}[l(y_*, \mu_1)] \geq E_{p(y_*|\vec{y})}[l(y_*, \mu_2)]$

Hence, Proved.

PROBLEM 2

(a) $\text{Gamma}(x | a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} = p(x | a, b)$.

$$-\log p(x | a, b) = -\log\left(\frac{b^a}{\Gamma(a)}\right) - (a-1)\log x + bx$$

By Laplace Approximation,

① $\rightarrow -\log p(x | a, b) \approx -\log p(\hat{x} | a, b) + \frac{(x - \hat{x})^2}{2\sigma^2}$, where $\sigma^{-2} = \left.\frac{\partial^2(-\log p(x|a,b))}{\partial x^2}\right|_{x=\hat{x}}$

and $\left.\frac{\partial(-\log p(x|a,b))}{\partial x}\right|_{x=\hat{x}} = 0$

Using these, we get $-\frac{\partial(\log p(x|a,b))}{\partial x} = -\frac{(a-1)}{x} + b = 0$. at $x=\hat{x}$.

$$\Rightarrow \hat{x} = \frac{a-1}{b}.$$

also, $\sigma^{-1} = \left.\frac{\partial(-\log p(x|a,b))}{\partial x^2}\right|_{x=\hat{x}} = \frac{a-1}{\hat{x}} = \frac{a-1}{\left(\frac{a-1}{b}\right)^2} = \frac{b^2}{a-1}$

$$\Rightarrow \sigma = \frac{a-1}{b^2}.$$

Taking exponent on both sides in equation ①, we get approximation $\tilde{p}(x|a,b) \approx e^{const.} \cdot e^{\frac{(x-\hat{x})^2}{2\sigma^2}}$

This is in the form of Normal distribution with mean $\mu = \hat{x} = \frac{a-1}{b}$ and variance $\sigma^2 = \frac{a-1}{b^2}$

Thus, $\tilde{p}(x|a,b) = \mathcal{N}\left(\frac{a-1}{b}, \frac{a-1}{b^2}\right) = \frac{b}{\sqrt{2\pi(a-1)}} \exp\left\{\frac{(bx-(a-1))^2}{2(a-1)}\right\}$.

Consider a gaussian approximation $\tilde{q}(x|a,b) = \mathcal{N}(x|\mu', \sigma^{2\prime})$ with μ' = mean of Gamma $(x|a,b) = \frac{a}{b}$.
and σ^2 = variance of Gamma $(x|a,b) = \frac{a}{b^2}$

Then $\tilde{q}(x|a,b) = \mathcal{N}\left(x \middle| \frac{a}{b}, \frac{a}{b^2}\right)$
and $\tilde{p}(x|a,b) = \mathcal{N}\left(x \middle| \frac{a-1}{b}, \frac{a-1}{b^2}\right)$
The $\tilde{q}(x|a,b)$ is obtained by setting a to $(a-1)$ in the Laplace approximation

Also,
$\tilde{q}(x|a,b)$ tends to $\tilde{p}(x|a,b)$ when $a \rightarrow \infty$, as the original distribution is flatter.

(b) Using Laplace approximation,

$$-\ln p(x|a,b) = -\ln p(\hat{x}|a,b) + \frac{(x-\hat{x})(x-\hat{x})}{2\sigma^2}, \quad \hat{x} = \frac{a-1}{b}, \quad \sigma^{-1} = \frac{b^2}{a-1}$$

putting $x = \hat{x}$ in above ~~equation~~ approximation,

$$\cancel{-\ln p(\hat{x}|a,b) = -\ln}$$

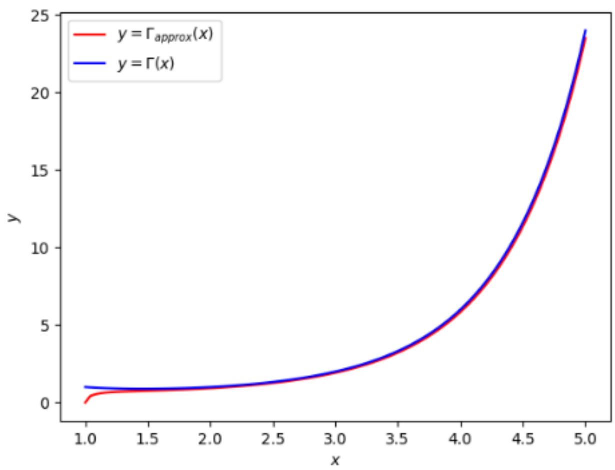
$$\text{Gamma}(\hat{x}|a,b) = \frac{b}{\sqrt{2\pi(a-1)}} \exp\left\{\frac{-(\hat{x}-\hat{x})^2}{2\sigma^2}\right\} = \frac{b}{\sqrt{2\pi(a-1)}}$$

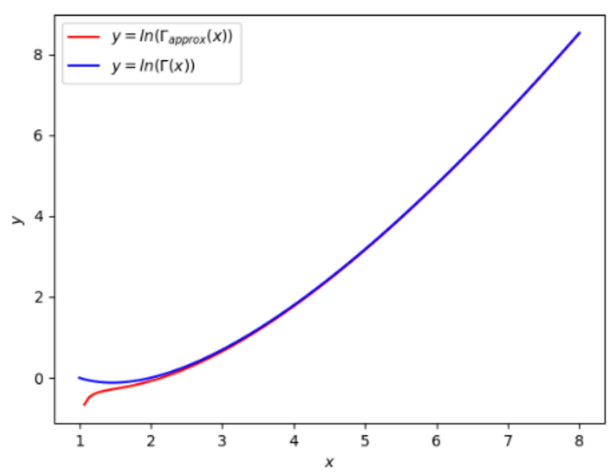
$$\frac{b^a}{\Gamma(a)} \left(\frac{a-1}{b}\right)^{a-1} e^{-b\left(\frac{a-1}{b}\right)} = \frac{b}{\sqrt{2\pi(a-1)}}$$

$$\Gamma(a) = \sqrt{2\pi}\,(a-1)^{a-1+1/2}\, b^{a-(a-1)-1}\, e^{-(a-1)}$$

$$\Gamma(a) = \sqrt{2\pi}\,(a-1)^{a-1/2}\, e^{-(a-1/2)+1/2}$$

$$\Gamma(a) = \sqrt{2\pi e}\left(\frac{a-1}{e}\right)^{a-1/2}$$





PROBLEM-3

Let single observation x be drawn from Gaussian,

$$x \sim \mathcal{N}(x|\mu, \beta^{-1})$$

with priors $\quad \mu \sim \mathcal{N}(\mu|\mu_0, s_0)$

and $\quad \beta \sim \text{Gamma}(\beta|a, b)$

Conditional distributions are $\underset{\text{treat } \beta \text{ as given}}{p(\mu|x,\beta)}$ and $\underset{\text{treat } \mu \text{ as given}}{p(\beta|x,\mu)}$

[ommitting hyperparameters from notation for simplicity]

$$p(\mu|x,\beta) \propto p(x|\mu,\beta)\,p(\mu|\mu_0, s_0)$$
$$\propto \mathcal{N}(x|\mu,\beta^{-1})\,\mathcal{N}(\mu|\mu_0, s_0)$$

[Since these are conjugate (locally due to β being treated as constant)]

We use the result that

$$\mathcal{N}(\mu_1, \beta_1^{-1}) \cdot \mathcal{N}(\mu_2, \beta_2^{-1}) \propto \mathcal{N}(\mu_3, \beta_3^{-1}) \text{ where } \beta_3 = \beta_1 + \beta_2$$
$$\text{and } \mu_3 = \frac{\beta_1}{\beta_3}\mu_1 + \frac{\beta_2}{\beta_3}\mu_2.$$

Using this, we obtain.

$$p(\mu|x,\beta) = \mathcal{N}(\mu|\mu', (\beta')^{-1}) \; ; \; \mu' = \frac{\mu_0}{\beta s_0 + 1} + \frac{\beta s_0 x}{\beta s_0 + 1}$$
$$\beta' = \beta + \frac{1}{s_0}$$

$$= \mathcal{N}\left(\mu \,\Big|\, \left(\frac{\mu_0}{\beta s_0 + 1}\right) + \left(\frac{\beta s_0 x}{\beta s_0 + 1}\right), \left(\beta + \frac{1}{s_0}\right)^{-1}\right)$$

Similarly, for β,

$$p(\beta|x,\mu) \propto p(x|\mu,\beta)\,p(\beta|a,b) \propto \mathcal{N}(x|\mu,\beta^{-1})\,\text{Gamma}(\beta|a,b)$$

$$\propto \frac{\beta^{1/2}}{\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2\beta}{2}\right\} \frac{b^a \beta^{a-1}}{\Gamma(a)} \exp\{-b\beta\}.$$

$$\propto \text{Gamma}(\beta|a', b') \text{ where } a' = a + 1/2$$
$$\text{and } b' = b + \frac{1}{2}(x-\mu)^2.$$

$$\Rightarrow p(\beta|x,\mu) = \text{Gamma}\left(\beta \,\Big|\, (a + 1/2), \left(b + \tfrac{1}{2}(x-\mu)^2\right)\right)$$

PROBLEM-4

The gamma-Normal prior is given as:

$$p(\mu, \tau \mid \mu_0, \lambda_0, \alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}\sqrt{\lambda_0}}{\Gamma(\alpha_0)\sqrt{2\pi}} \tau^{(\alpha_0 - 1/2)} \exp\left\{-\beta\tau - \frac{\lambda_0 \tau(\mu - \mu_0)^2}{2}\right\}.$$

with parameters as $\mu_0, \lambda_0, \alpha_0$ and β_0

Consider exponential family form as

$$p(x \mid \theta) = h(x)\exp\left\{\theta^T \phi(x) - A(\theta)\right\};$$

θ are the natural parameters.
$\phi(x)$ is the sufficient statistics
$A(\theta)$ is the log partition function.
$h(x)$ is a function of x (usually a constant).

converting the gamma-Normal function to exponential family form:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\beta_0\tau - \frac{\lambda_0\tau\mu^2}{2} + \lambda_0\tau\mu\mu_0 - \frac{\lambda_0\tau\mu_0^2}{2} + \ln\left(\frac{\beta_0^{\alpha_0}\sqrt{\lambda_0}}{\Gamma(\alpha_0)}\tau^{(\alpha_0 - 1/2)}\right)\right\}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{\left(-\beta_0\tau - \frac{\lambda_0\tau\mu^2}{2} + \lambda_0\tau\mu\mu_0 - \frac{\lambda_0\tau\mu_0^2}{2} + (\alpha_0 - 1/2)\ln\tau\right) - \left(-\alpha_0\ln\beta_0 - \tfrac{1}{2}\ln\lambda_0 + \ln\Gamma(\alpha_0)\right)\right\}$$

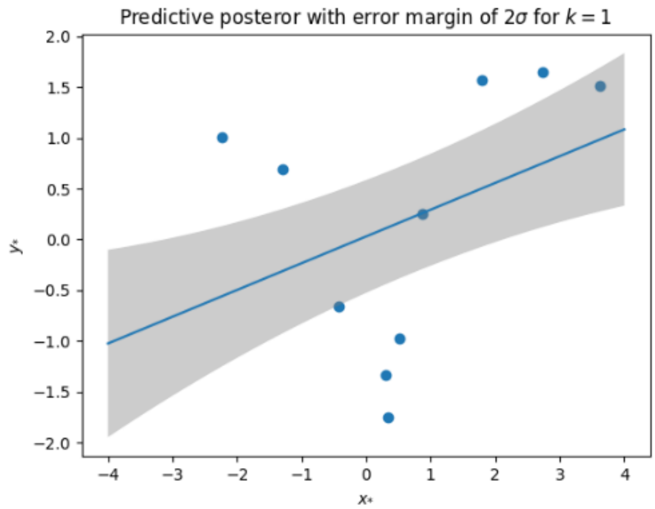
$$= \frac{1}{\sqrt{2\pi}} \exp\left\{\begin{bmatrix}\alpha_0 - 1/2 \\ -\beta_0 - \frac{\lambda_0\mu_0^2}{2} \\ \lambda_0\mu_0 \\ -\lambda_0/2\end{bmatrix}^T \begin{bmatrix}\ln\tau \\ \tau \\ \tau\mu \\ \tau\mu^2\end{bmatrix} - \left(\ln\Gamma(\alpha_0) - \tfrac{1}{2}\ln\lambda_0 - \alpha_0\ln\beta_0\right)\right\}.$$

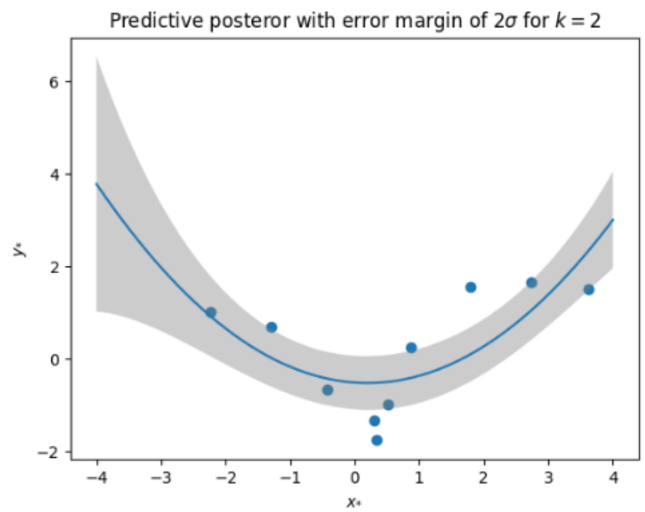
Comparing the 2 forms, we observe that.

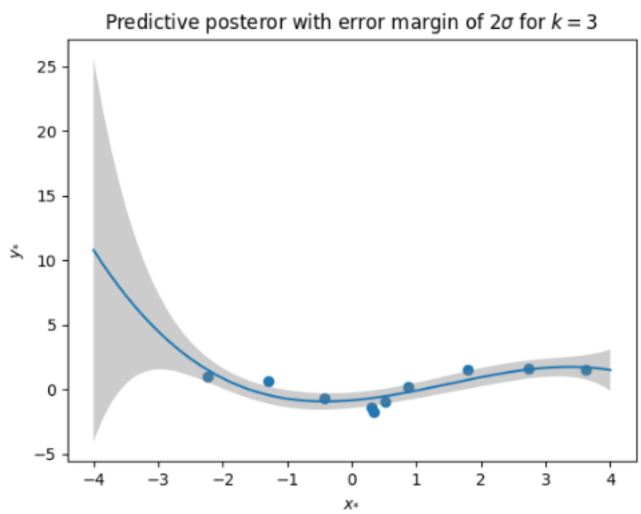
$$\theta = \begin{bmatrix}\alpha_0 - 1/2 \\ -\beta_0 - \frac{\lambda_0\mu_0^2}{2} \\ \lambda_0\mu_0 \\ -\lambda_0/2\end{bmatrix}$$ natural parameters; $$\phi(x) = \begin{bmatrix}\ln\tau \\ \tau \\ \tau\mu \\ \tau\mu^2\end{bmatrix}$$ sufficient statistics

$h(x) = \frac{1}{\sqrt{2\pi}}$, a constant ; $A(\theta) = \log Z(\theta) = \ln\Gamma(\alpha_0) - \tfrac{1}{2}\ln\lambda_0 - \alpha_0\ln(\beta_0)$

Problem 5


Predictive posteror with error margin of 2σ for k = 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y*
−4
−3
−2
−1
0
1
2
3
4
x*


Predictive posteror with error margin of 2σ for k = 2
6
4
2
0
−2
y*
−4
−3
−2
−1
0
1
2
3
4
x*


Predictive posteror with error margin of 2σ for k = 3
25
20
15
10
5
0
−5
y*
−4
−3
−2
−1
0
1
2
3
4
x*

Marginal likelihood is $p(\vec{y} \mid \phi_k(\vec{x}), \beta)$ for $k=1,2,3$.

On computing, we get

$k=1$, marginal likelihood = 1.49e-08

$k=2$, marginal likelihood = 5.10e-09

$k=3$, marginal likelihood = 3.92e-10

From the plots, it seems that model with $k=3$ explains data the best. But from marginal likelihood, $k=3$ is least. This is due to more degrees of freedom.

Given an additional training input (x', y'), we should choose x' to be in the region where there is highest variance, or lowest concentration of data points.

For $k=1$, x' should be around -4 to -3 or 3 to 4.

for $k=2$, x' should be around -4 to -3.

for $k=3$, x' should be around -4.

(a)

Link for the code: https://goo.gl/eJUD6A