(1) As we know: $\Gamma(2,1)$ PDF : $P(\lambda) = \lambda \exp(-\lambda)$, $\lambda > 0$.

$$P_X(x) = \frac{\lambda^x}{x!} \exp(-\lambda), \quad \lambda > 0, \quad x = 0, 1, 2 \ldots$$

because $x_1, x_2 \ldots x_d$ is an iid

MAP: $\arg\max_{\lambda} P(\omega|x) = \arg\max_{\lambda} \frac{P(x|\lambda) P(\lambda)}{P(x)}$.

$$\propto \arg\max_{\lambda} P(x|\lambda) P(\lambda).$$

$$= \arg\max_{\lambda} \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \right) \cdot \lambda \exp(-\lambda)$$

$$= \arg\max_{\lambda} \sum_{i=1}^{n} [x_i \ln\lambda - \lambda - \ln x_i! + \ln\lambda - \lambda$$

$$= \arg\max_{\lambda} [\ln\lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i!) + \ln\lambda - \lambda] \to 1$$

$$\frac{\partial \theta}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\lambda. \to \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i > 0.$$

$$\frac{\partial^2 \theta}{\partial \lambda^2} = \frac{-\sum_{i=1}^{n} x_i}{\lambda^2} < 0.$$

$\therefore \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i$. is a value that maximizes the posterior.

2. $P_{X|\lambda}(X|\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda) \quad x = 0, 1, 2 \cdots$

$L(\lambda|D) = \prod_{i=1}^{n} P(x_i|\lambda) = \sum_{i=1}^{n} \ln P(x_i|\lambda)$

$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{n} [x_i \ln\lambda - \lambda - \ln(x_i!)]$

set $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

$= \arg\max_{\lambda} [\ln(\lambda) \cdot n\mu - n\lambda - \sum_{i=1}^{n} \ln(x_i!)] \quad \cdots \quad ①$

$\frac{\partial \theta}{\partial \lambda} = \frac{n\mu}{\lambda} - n = 0 \Rightarrow \lambda = \mu.$

$\frac{\partial^2 \theta}{\partial \lambda^2} = -\frac{n\mu}{\lambda^2} = -\frac{n}{\mu} < 0. \leftarrow$ maxium.

$\Rightarrow \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$

CLT: given enough samples, data will follow normal distribution.

$N(\mu, \sigma^2)$. which means

$\hat{\lambda} = \arg\max_{\lambda} N(\mu, \sigma^2) = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i.$

From CLT, $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$

(3)

$$Y = \beta_0 + \beta_1 x + \cdots \beta_p x_p + \varepsilon \quad , \quad \varepsilon \sim N(0, 6^2)$$

$\Rightarrow Y = x\beta + \varepsilon$, where $x = [1, x_1, \ldots x_p]$

As we know, least squares: $\beta = (x^T x)^{-1} x^T y$.

MLE : $\hat{\beta} = \underset{\beta}{\arg\max}\ P(Y|D) \quad D = \{x_1 \cdots x_p\}$.

$$= \underset{\beta}{\arg\max} \prod_{i=1}^{n} P(Y|x_i)$$

$\varepsilon = y - x\beta = N(0, 6^2)$

$$\hat{\beta} = \underset{\beta}{\arg\max} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\, 6^2} \exp\left(-\frac{1}{2}\left(\frac{y_i - x\beta_i}{6}\right)^2\right).$$

$$= \underset{\beta}{\arg\max} \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi}\, 6^2} - \frac{1}{2}\left(\frac{y_i - x\beta_i}{6}\right)^2 \quad \cdots\cdots \ ①$$

$$\frac{\partial ①}{\partial \beta} = \sum_{i=1}^{n} \frac{x^T}{6}\left(\frac{y_i - x\beta_i}{6}\right) = 0 . \Rightarrow x^T x \beta = x^T y .$$

$\therefore \hat{\beta} = (x^T x)^{-1} x^T y$

$$\frac{\partial^2 ①}{\partial \beta^2} = \frac{-x^T x}{6^2} < 0 . \leftarrow \text{maximum.}$$

(4) $Y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon$. , $\varepsilon \sim N(0, \sigma^2)$.

$Y = x\hat{\beta} + \varepsilon \qquad x = [1, x_1, \cdots x_p]$

$\varepsilon = y - x\hat{\beta} \sim N(0, \sigma^2)$

$P_{Y|X}(Y|x_i) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp(-\frac{1}{2}(\frac{y - x_i\beta_i}{\sigma})^2)$.

MAP: $\hat{\beta} = \arg\max_{\beta} [\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{y - x_i\beta_i}{\sigma})^2)] \cdot \sqrt{\frac{\lambda}{2\pi\sigma^2}} \exp(-\frac{\lambda}{2}(\frac{\beta_i}{\sigma})^2]$.

$\Rightarrow \hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{p} [-\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{1}{2}(\frac{y - x_i\beta_i}{\sigma})^2] + \frac{1}{2}\ln\lambda - \frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{\lambda}{2}(\frac{\beta_i}{\sigma})^2$

$= \arg\max [\frac{p+1}{2}\ln 2\pi + (p+1)\ln\sigma) - \frac{1}{2}\ln(\lambda) + \frac{\lambda}{2}(\frac{\beta_i}{\sigma})^2 + \frac{1}{2}(\frac{y - x_i\beta_i}{\sigma})^2\} \cdots ①$

$\frac{\partial ①}{\partial \beta} = \frac{\lambda\beta_i}{\sigma^2} - x^T(\frac{y - x\beta}{\sigma^2}) = 0 \Rightarrow (x^Tx + \lambda)\beta = x^Ty$.

$\Rightarrow \hat{\beta} = (x^Tx + \lambda)^{-1} x^Ty$.

Based on probabalistic structure, λ should add bias to estimate.

$\Rightarrow \hat{\beta} = (x^Tx + \lambda\Sigma)^{-1} x^Ty$.

(t) $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, 6^2)$.

$P(\beta_i) = lap(0, \frac{6^2}{\lambda}) = \frac{\lambda}{26^2} \exp(\frac{-|\beta_i|}{6^2})$

$\varepsilon = y - X\beta \sim N(0, 6^2)$

$P_{Y|X}(Y|X_i) = \frac{1}{\sqrt{2\pi 6^2}} \exp(-\frac{1}{2}(\frac{(y - X_i\beta_i)^2}{6}))$

MAP: $\hat{\beta} = \arg\max_{\beta} [\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}6} \exp(-\frac{1}{2}(\frac{y - X_i\beta_i}{6})^2)] \cdot \frac{\lambda}{26^2} \exp(-\frac{|\beta_i|}{6^2})$

$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{p} [\frac{1}{2}\ln(2\pi) + \ln(6) + \frac{1}{2}(\frac{y - X_i\beta_i}{6})^2 + \ln(\frac{\lambda}{26^2}) + \frac{|\beta_i|}{6^2}$

$$\nabla_\beta \hat{\beta} = \begin{cases} \frac{1}{6^2}[-X^T(Y - X\beta) + 1] & \beta_i \geq 0. \\ \frac{1}{6^2}[-X^T(Y - X\beta) - 1] & \text{otherwise}. \end{cases} = 0$$

$\beta_i \geq 0$: $(X^TX)\beta = X^TY - \frac{1}{6^2} \rightarrow \beta = (X^TX)^{-1}(X^TY - \frac{1}{6^2})$

$\beta_i < 0$: $(X^TX)\beta = X^TY + \frac{1}{6^2} \rightarrow \beta = (X^TX)^{-1}(X^TY + \frac{1}{6^2})$

$\nabla^2_\beta \beta = X^TX > 0 \rightarrow$ positive.

$\Rightarrow$ therefore, it's maxium.

$$\hat{\beta} = \begin{cases} (X^TX)^{-1}(X^TY - \frac{1}{6^2}) & \beta_i > 0. \\ (X^TX)^{-1}(X^TY + \frac{1}{6^2}) & \text{otherwise}. \end{cases}$$

(b) As we know $X = U\Sigma V^T$

a. $\hat{\beta}^{Ridge} = \underset{\beta}{\arg\min}\ RSS + \lambda \sum_{j=1}^{p} \beta_j^2 = \underset{\beta}{\arg\min}\ \|y - X\beta\|^2 + \lambda\|\beta\|^2.$

$= \underset{\beta}{\arg\min}\ (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$

$= \underset{\beta}{\arg\min}\ (y^Ty - (X\beta)^Ty + y^TX\beta + \beta^T(X^TX + \lambda I)\beta) \Leftarrow ①$

$\frac{\partial ①}{\partial \beta} = -X^Ty - (y^TX)^T + 2\beta(X^TX + \lambda I) = 0.$

$\Rightarrow \beta(X^TX + \lambda I) = X^Ty.$

$\Rightarrow \hat{\beta}^{Ridge} = (X^TX + \lambda I)^{-1}X^Ty$

$\Rightarrow \hat{\beta}^{Ridge} = ((U\Sigma V^T)^T(U\Sigma V^T) + \lambda I)^{-1}(V\Sigma^TU^T)y.$

$= (V\Sigma^TU^TU\Sigma V^T + \lambda I)^{-1}(V\Sigma^TU^T)y$

$\because \Sigma^T = \Sigma$

$\Rightarrow \hat{\beta}^{Ridge} = (\Sigma^T\Sigma + \lambda I)^{-1}(V\Sigma U^T)y.$

$\hat{y} - X\hat{\beta}^{Ridge} = U\Sigma V^T(\Sigma^T\Sigma + \lambda I)^{-1}(V\Sigma U^T)y$

$= \sum_{j=1}^{p} \sigma_j u_j v_j^T(\sigma_j^2 + \lambda)^{-1}v_j \sigma_j u_j^T y.$

$= \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j v_j^T v_j u_j^T y.$

$= \sum_{j=1}^{p} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T y.$

(b) $$X(X^TX+\lambda I)^{-1}X^T = U\Sigma V^T(V\Sigma^T U^T U\Sigma V^T + \lambda I)^{-1} V\Sigma U^T$$

$$= U\Sigma V^T(V\Sigma^T\Sigma V^T+\lambda I)^{-1} V^T\Sigma U$$

$$= U\Sigma V^T(\Sigma^T\Sigma(VV^T)+\lambda I)^{-1} V\Sigma U^T$$

$$= \sum_{i=1}^{p} u_i \frac{\sigma_i^2}{\sigma_i^2+\lambda} u_i^T$$

$$\Rightarrow tr\left(\sum_{i=1}^{p} u_i\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right) u_i^T\right) = \sum_{j=1}^{n}\sum_{i=1}^{p}\left(u_i \frac{\sigma_i^2}{\sigma_i^2+\lambda} u_i^T\right)_{jj}.$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{p}\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda} u_i u_i^T\right)_{jj}$$

$$= \sum_{i=1}^{n}\frac{\sigma_i^2}{\sigma_i^2+\lambda}$$

$$\therefore tr(X(X^TX+\lambda I)^{-1}X^T) = \sum_{i=1}^{n}\frac{\sigma_i^2}{\sigma_i^2+\lambda}$$