① PDF of $\Gamma(2,1)$: $P(z) = z\exp(-z),\ z>0$

Pmf of possion random variable: $P_X(x) = \lambda^x e^{-\lambda}/x!$

$$\arg\max_{\lambda} P(w|x) = \arg\max_{\lambda} \frac{P(x|z)P(z)}{P(x)}$$

$$= \arg\max_{\lambda} p(x|z)\,p(z)$$

$$= \arg\max_{\lambda} \prod_{i=1}^{n} (\lambda^{x_i}\exp(-\lambda)/x!)\, z\exp(-z)$$

$$\overset{\text{iid}}{=} \arg\max_{\lambda} \sum_{i=1}^{n} x_i \ln(\lambda) - \lambda - \ln(x_i!)] + \ln(z) - z$$

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

$$\frac{\partial}{\partial\lambda}\left(\arg\max_{\lambda}\left[\ln(\lambda)\sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n}\ln(x_i!) + \ln(z) - z\right]\right)$$

$$= \frac{ny}{\lambda} - n = 0$$

$$\Rightarrow y = \lambda > 0$$

$$\frac{\partial^2}{\partial\lambda^2}\left(\arg\max\left[\ln(\lambda)\sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n}\ln(x_i!) + \ln(z) - z\right]\right) = \frac{-ny}{\lambda^2}$$

$$= \frac{-ny}{y^2} = \frac{-n}{y} < 0$$

from sign of the second derivative. we can know that λ is the value that maximizes the poterior probability

② $P_X(x) = \lambda^x e^{-\lambda} / x!$ $(\lambda > 0, x = 0, 1 \cdots)$

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

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

$= \underset{\lambda}{\operatorname{argmax}} [\ln(\lambda) n y - n\lambda - \Sigma \ln(x_i)!]$

$\frac{\partial}{\partial \lambda} (\underset{\lambda}{\operatorname{argmax}} [\ln(\lambda) n y - n\lambda - \Sigma \ln(x_i!)]) = \frac{ny}{n} - n = 0$

$\hat{\lambda} = y$

$\frac{\partial^2}{\partial \lambda} (\underset{\lambda}{\operatorname{argmax}} [\ln(\lambda) n y - n\lambda - \Sigma \ln(x_i!)]) = \frac{-n\mu}{\lambda^2} = \frac{-n}{y} < 0$

$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$ (from fisher), CLT: data would follow $N(\mu, \sigma^2)$

$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \; N(\mu, \sigma^2) = \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$

from CLT we can get, $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$

③ $Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + e$ where $\varepsilon \sim N(0, \sigma^2)$

we can simplify as $Y = X\beta + \varepsilon$ where $X = [1, X_1 \dots X_p]$

$\beta = (X^T X)^{-1} X^T y$

MLE: $\hat{\beta} = \arg\max_{\beta} P(Y|D)$, $D = \{X_1, \dots, X_p\}$

$= \arg\max_{\beta} \prod_{i=1}^{n} P(Y|X_i)$

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

$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - x\beta_i}{\sigma}\right)^2} = \underbrace{\arg\max_{\beta} \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\left(\frac{y_i - x\beta_i}{\sigma}\right)^2\right]}_{①}$

$$\frac{\partial ①}{\partial \beta} = \sum_{i=1}^{n} \frac{X_i}{\sigma}\left(\frac{y_i - X_i\beta_i}{\sigma}\right) = 0$$

$$= \frac{X^T}{\sigma^2}(y - X\beta) = 0$$

$\Rightarrow X^T X\beta = X^T y \Rightarrow \beta = (X^T X)^{-1} X^T y$

$$\frac{\partial^2 ①}{\partial \beta^2} = \frac{-X^T X}{\sigma^2} < 0$$

the negative sign makes maximizing the log-likelihood equivalent to minizing the negative log-likelihood

$\therefore$ β vector obtained by least square and maximum likelihood are the same

4. $Y = \beta_0 + \beta_1 X_1 + \dots \beta_p x_p + e$ where $\varepsilon \sim N(0, \sigma^2)$

We can simplify as $Y = x\beta + \varepsilon$ where $X = [1, x_1 \dots x_p]$

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

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

$$MAP = \hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left[\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y - x_i\beta_i}{\sigma}\right)^2\right]\right] \cdot$$

$$\sqrt{\frac{\lambda}{2\pi\sigma^2}} \exp\left[-\frac{\lambda}{2}\left(\frac{\beta_i}{\sigma}\right)^2\right]$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{p} \left[-\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2}\left(\frac{y - \beta_i x_i}{\sigma}\right)^2\right] +$$

$$\frac{1}{2}\ln(\lambda) - \frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{\lambda}{2}\left(\frac{\beta_i}{\sigma}\right)^2$$

$$= \underset{\lambda}{\operatorname{argmax}} \left[\frac{p+1}{2}\ln(2\pi) + (p+1)\ln(\sigma) - \frac{1}{2}\ln(\lambda) + \frac{\lambda}{2}\left(\frac{\beta_i}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{y - x_i\beta_i}{\sigma}\right)^2\right] \quad (1)$$

$$\frac{\partial (1)}{\partial \beta} = \frac{\lambda\beta_i}{\sigma^2} - x^T\left(\frac{y - x\beta}{\sigma^2}\right) = 0$$

$$(x^T x + \lambda)\beta = x^T y$$

$$\hat{\beta} = (x^T x + \lambda I)^{-1} x^T y$$

④ $P(\beta \mid Y, X_1, \dots X_p) \propto P(Y \mid \beta, X_1 \dots X_p) \cdot P(\beta)$

$$\Rightarrow P(Y \mid X_1, \dots X_p) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y-\beta_0-\beta_1 X_1 \dots \beta_p X_p)^2 / 2\sigma^2}$$

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

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

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

$$(Y - X\beta = \quad X = \{1, X_1, \dots X_p\})$$

case 1: $\beta_i \geq 0$

$$(X^TX)\beta = X^TY - \frac{1}{\sigma^2}$$

$$\beta = (X^TX)^{-1}(X^Ty - \frac{1}{\sigma^2})$$

case 2: $\beta_i < 0$

$$(X^TX)\beta = X^TY + \frac{1}{\sigma^2}$$

$$\beta = (X^TX)^{-1}(X^TY + \frac{1}{\sigma^2})$$

$$\nabla^2_\beta \hat{\beta} = X^TX > 0$$

$$\therefore \hat{\beta} = \begin{cases} (X^TX)^{-1}(X^TY - \frac{1}{\sigma^2}) & \beta_i \geq 0 \\ (X^TX)^{-1}(X^TY + \frac{1}{\sigma^2}) & \beta_i < 0 \end{cases}$$

$\therefore \vec{\beta}$ is no longer an unbiased estimator

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

(a) Show: $\vec{y} = X\hat{\beta}^{Ridge} = \sum_{j=1}^{p} u_j \frac{\sigma_j^2}{\sigma_j^2+\lambda} u_j^T y$

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

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

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

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

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

$$2\beta(X^TX + \lambda I) = 2X^Ty$$

$$\vec{\beta}^R = (X^TX + \lambda I)^{-1}X^Ty$$

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

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

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

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

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

$$\therefore \hat{y} = X\hat{\beta}^R = \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2+\lambda} u_j u_j^T y$$

(b) Show $tr[X(X^TX+\lambda I)^{-1}X^T] = \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2+\lambda}$ using SVD

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

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

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

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

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

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

$$= \sum_{j=1}^{n} \frac{\sigma_j^2}{\sigma_j^2+\lambda}$$

$$\therefore trace(X(X^TX+\lambda I)^{-1}X^T) = \sum_{j=1}^{n} \frac{\sigma_j^2}{\sigma_j^2+\lambda}$$