HW 4 664851 7090 Yu Wei Chen

1.

$\Gamma(2,1)$ PDF : $p(z) = z \exp(-z)$, $z < 0$

Possion PMF : $p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $\lambda > 0$, $x = 0, 1, \ldots$

$$\text{MAP} = \underset{\lambda}{\arg\max}\ P(w|x)$$

$$= \underset{\lambda}{\arg\max} \frac{p(x|z)\,p(z)}{p(x)}$$

$$= \underset{\lambda}{\arg\max}\ p(X|Z)\,p(Z)$$

$$= \underset{\lambda}{\arg\max} \prod_{i=1}^{n} \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right) z e^{(-z)}$$

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

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

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

$$= \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0 \Rightarrow \sum_{i=1}^{n} x_i = \lambda n \rightarrow \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i > 0$$

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

$$= \frac{-\sum_{i=1}^{n} x_i}{\lambda^2} < 0 \qquad \therefore \lambda_{MAP} = \frac{\sum_{i=1}^{n} x_i + 1}{n+1}$$ is a value that maximizes the posterior QED ∎

2.

$$P_{X|\lambda}(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda > 0, \quad x = 0, 1, \dots$$

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

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

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

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

$$= \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0 \quad \Rightarrow \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial^2 \underset{\lambda}{\operatorname{argmax}} \left[\ln(\lambda) \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i!) \right]}{\partial \lambda^2}$$

$$= \frac{-\sum_{i=1}^{n} x_i}{\lambda^2} = \frac{n^2}{\sum_{i=1}^{n} x_i} < 0 \rightarrow \text{Maximum}$$

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

CLT: Given enough samples, data will follow normal distribution $N(\frac{1}{n}\sum_{i=1}^{n} x_i, \sigma^2)$. This means that $\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} N(\frac{1}{n}\sum_{i=1}^{n} x_i, \sigma^2) = \frac{1}{n}\sum_{i=1}^{n} x_i$

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

3.

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

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

Least square estimate: $\beta = (X^T X)^{-1} X^T y$

MLE: $\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ p(Y|D)$. $D = \{X_1, \dots, X_p\}$

$$= \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{n} p(Y|X_i)$$

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

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y - X\beta_i)^2}{2\sigma^2}\right)$$

$$= \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(Y - X\beta_i)^2}{2\sigma^2} \right]$$

$$\frac{\partial \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(Y - X\beta_i)^2}{2\sigma^2} \right]}{\partial \beta}$$

$$= \sum_{i=1}^{n} \frac{X^T}{\sigma^2}(Y - X\beta_i)$$

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

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

$$\frac{\partial^2 \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(Y - X\beta)^2}{2\sigma^2} \right]}{\partial \beta^2}$$

$$= \frac{-X^T X}{\sigma^2} < 0 \quad \rightarrow \quad \text{Maximum.}$$

The MLE and least square estimate are the same, and by the Gauss - Markov Theorem, the estimator is also BLUE. $\blacksquare$

4.

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

$$Y = X\beta + \varepsilon \quad . \quad X = [1, X_1, \dots, X_p]$$

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

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

$$\text{MAP}: \hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ p(\beta|Y)$$

$$= \underset{\beta}{\operatorname{argmax}}\ p(Y|\beta)\cdot p(\beta)$$

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

$$= \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n}\left[-\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2\sigma^2}(y_i - x_i\beta)^2\right]$$

$$+ \sum_{i=1}^{p} \frac{1}{2}\ln(\lambda) - \frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{\lambda}{2\sigma^2}(\beta_i)^2$$

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

$$\frac{\partial\ \underset{\beta}{\operatorname{argmin}} \left[\frac{n+p}{2}\ln(2\pi) + (n+p)\ln(\sigma) - \frac{p}{2}\ln(\lambda) + \frac{\lambda}{2\sigma^2}(\beta)^2 + \frac{1}{2\sigma^2}(y - x\beta)^2\right]}{\partial\beta}$$

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

$$\lambda\beta - x^Ty + x^TX\beta = 0$$

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

λ adds bias to estimate to give a more accurate estimated based on probabilistic structure.

5.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$Y = X\beta + \varepsilon$$

$$\beta_i \sim Lap(0, \frac{\sigma^2}{\lambda})$$

$$p(\beta_i) = \frac{\lambda}{2\sigma^2} \exp\left(-\frac{\lambda|\beta_i|}{\sigma^2}\right)$$

$$p(\beta) = \prod_{i=1}^{p} \frac{\lambda}{2\sigma^2} \exp\left(-\frac{\lambda}{\sigma^2}|\beta_i|\right)$$

$$p(Y|\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right)$$

$$p(\beta|Y) \propto p(Y|\beta)\, p(\beta)$$

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}} \left[p(Y|\beta) \cdot p(\beta) \right]$$

$$= \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2\sigma^2}(y_i - x_i\beta)^2 \right]$$

$$+ \sum_{i=1}^{p} \left[\ln(\lambda) - \ln(2\sigma^2) - \frac{\lambda}{\sigma^2}|\beta_i| \right]$$

$$= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\frac{1}{2}\ln(2\pi) + \ln(\sigma) + \frac{1}{2\sigma^2}(y_i - x_i\beta)^2 \right] + \sum_{i=1}^{p} \left[\ln(2\sigma^2) - \ln(\lambda) + \frac{\lambda}{\sigma^2}|\beta_i| \right]$$

$$= \underset{\beta}{\operatorname{argmin}} \left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \frac{\lambda}{\sigma^2} \sum_{i=1}^{p} |\beta_i| \right]$$

$$= \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - x_i\beta)^2 + 2\lambda \sum_{i=1}^{p} |\beta_i| \right]$$

The MAP estimate corresponds to solving a LASSO regression problem, where the L_1 penalty comes from the Laplace prior. This leads to sparse solutions where some β_i are exactly zero. The parameter λ controls the strength of the shrinkage and the level of sparsity.

6.
(a)

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

$$X = U \Sigma V^T$$

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

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

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

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

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

$$\Sigma^2 + \lambda I = \mathrm{diag}(\sigma_1^2 + \lambda, \sigma_2^2 + \lambda, \cdots, \sigma_p^2 + \lambda)$$

$$\Sigma (\Sigma^T \Sigma + \lambda I)^{-1} \Sigma = \mathrm{diag}\left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right)$$

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

(b)

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

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

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

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

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

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

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

$$= \sum_{j=1}^{n} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \underbrace{\sum_{i=1}^{p} (u_i u_i^T)_{jj}}_{I}$$

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

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