P1

a) $P(X|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$

$L(\lambda; x_1, \dots x_n) = \prod_{j=1}^{n} e^{\lambda} \frac{1}{x_j!} \lambda^{x_j}$

b)

$L(\lambda; x_1, \dots x_n) = -n\lambda - \sum_{j=1}^{n} (\ln(x_j!)) + (\ln(\lambda)) \sum_{j=1}^{n} x_j$

$\frac{\partial L(\lambda; x_1, \dots x_n)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{j=1}^{n} x_j = 0$

$\lambda_{MLE} = \frac{1}{n} \sum_{j=1}^{n} x_j$

c) let $\bar{x} = (x_1, \dots, x_n)$ be a vector.

$P(\lambda) = \text{gamma}(a, b)$ as a prior

$\lambda \mapsto \frac{f(\bar{x}; \lambda) P(\lambda)}{\int_{\lambda} f(x; \lambda) P(\lambda) \, d\lambda}$

$\lambda_{MAP} = \text{argmax} \frac{f(x|\lambda) P(\lambda)}{\int f(x|\lambda) P(\lambda) d\lambda} = \underset{\lambda}{\text{argmax}} \, f(x|\lambda) P(\lambda)$

$\lambda_{MAP} = \underset{\lambda}{\text{argmax}} \, P(\lambda) \prod_{i=1}^{n} P(x_i|\lambda) = \underset{\lambda}{\text{argmax}} \, (\log(P(\lambda)) + \sum_{i=1}^{n} \log f(x_i|\lambda))$

$\log f(x_i|\lambda) = x_i \log(\lambda) - \lambda$

$\sum_{i=1}^{n} \log f(x_i|\lambda) = \log\lambda \sum_{i=1}^{n} x_i - n\lambda$

$\frac{\partial \sum_{i=1}^{n} \log f(x_i|\lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$

$\log P(\lambda) = \log\left[\frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)\right]$

$= \log \frac{b^a}{\Gamma(a)} + (a-1)\log(\lambda) - b\lambda$

$\frac{\partial \log P(\lambda)}{\partial \lambda} = \frac{a-1}{\lambda} - b$

$\frac{\partial (\log(P(\lambda)) + \sum_{i=1}^{n} \log f(x_i|\lambda))}{\partial \lambda} = \frac{a-1}{\lambda} - b + \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$

$\frac{a-1+\sum x_i}{\lambda} = n + b$

$\lambda_{MAP} = \frac{a-1+\sum_{i=1}^{n} x_i}{n+b}$

d)

From Bayes rules,

$$P(\lambda|\bar{X}) = \frac{P(\bar{X}|\lambda)\,P(\lambda)}{\int_0^1 P(\bar{X}|\lambda)\,P(\lambda)\,d\lambda}$$

$$P(\lambda|\bar{X}) \propto P(\bar{X}|\lambda)\,P(\lambda)$$

$$P(\bar{X}|\lambda)\,P(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{1}{x_i!}\lambda^{x_i}\,\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

$$P(\lambda|\bar{X}) \propto \prod_{i=1}^{n}(\lambda^{x_i}e^{-\lambda})\,\lambda^{a-1}e^{-b\lambda}$$

$$= \lambda^{\sum_{i=1}^{n}x_i + a - 1}\,e^{-(n+b)\lambda}$$

which is proportional to pdf of gamma$(\sum_{i=1}^{n}x_i + a,\ n+b)$

Posterior dist of λ should be gamma$(\sum_{i=1}^{n}x_i + a,\ n+b)$

e) $\lambda|\bar{X} \sim$ gamma$(\sum_{i=1}^{n}x_i + a,\ n+b)$

$$E(\lambda|\bar{X}) = \frac{\sum_{i=1}^{n}x_i + a}{n+b}$$

$$\lambda_{ML} = \frac{\sum x_i}{n}$$

$$\lambda_{MAP} = \frac{\sum x_i + a - 1}{n+b}$$

$$\text{Var}(\lambda|\bar{X}) = \frac{\sum_{i=1}^{n}x_i + a}{(n+b)^2}$$

mean of λ under the posterior is close to λ_{ML} and λ_{MAP}. In particular, it's only off by $\frac{1}{n+b}$ from the λ_{MAP}. The mean of λ under posterior can also be expressed by a compromise between prior mean $\frac{a}{b}$ and the λ_{ML}.

$$\frac{\sum x_i + a}{n+b} = \lambda_{ML}\left(1 - \frac{b}{n+b}\right) + \frac{a}{b}\left(\frac{b}{n+b}\right)$$

As $n = 0$, mean reduces to $\frac{a}{b}$ prior mean

As $n \to \infty$, mean reduces to λ_{ML}

Problem 2

a) $y_i \overset{iid}{\sim} N(x_i^T w, \sigma^2)$

$w_{RR} = (\lambda I + X^T X)^{-1} X^T y$

$E(w_{RR}) = E[(\lambda I + X^T X)^{-1} X^T y]$

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

$= (\lambda I + X^T X)^{-1} X^T X w$

$Var(w_{RR}) = E[(w_{RR} - E(w_{RR}))(w_{RR} - E(w_{RR}))^T]$

$= E[w_{RR} w_{RR}^T] - E(w_{RR}) E(w_{RR})^T$

$Var(w_{RR}) = E[(X^T X + \lambda I)^{-1} X^T y y^T X (X^T X + \lambda I)^{-1}] - E(w_{RR}) E(w_{RR})^T$

$= (X^T X + \lambda I)^{-1} X^T E(yy^T) X (X^T X + \lambda I)^{-1} - E(w_{RR}) E(w_{RR})^T$

$Var(w_{RR}) = (X^T X + \lambda I)^{-1} X^T (\sigma^2 I + X w w^T X^T) X (X^T X + \lambda I)^{-1} - E(w_{RR}) E(w_{RR})^T$

$E(w_{RR}) E(w_{RR})^T = (X^T X + \lambda I)^{-1} X^T X w w^T X^T X (X^T X + \lambda I)^{-1}$

$Var(w_{RR}) = (X^T X + \lambda I)^{-1} X^T \sigma^2 I X (X^T X + \lambda I)^{-1}$

let $Z = (X^T X + \lambda I)^{-1}$

$Var(w_{RR}) = Z X^T \sigma^2 I X Z = \sigma^2 Z X^T X Z$

b)

$w_{RR} = (X^T X + \lambda I)^{-1} X^T y = \underbrace{(X^T X + \lambda I)^{-1} X^T X}_{S} w_{LS}$

$= S w_{LS}$

let $X = USV^T$ $\quad (X^TX)^{-1} = VS^{-2}V^T$

$$w_{RR} = (\lambda(X^TX)^{-1} + I)^{-1} w_{LS}$$

$$= (\lambda VS^{-2}V^T + I)^{-1} w_{LS}$$

$$= V(\lambda S^{-2} + I)^{-1} V^T w_{LS}$$

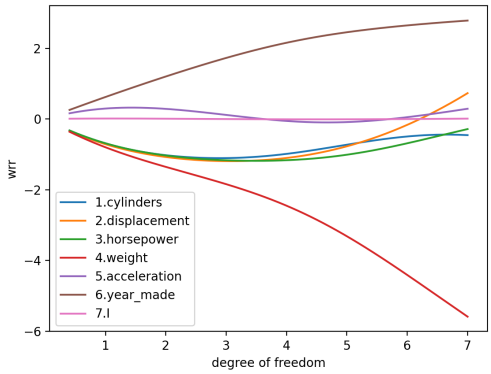
$$= VMV^T w_{LS}$$

with M to be a diagonal matrix with $M_{ii} = \frac{s_i^2}{\lambda + s_i^2}$

Problem3

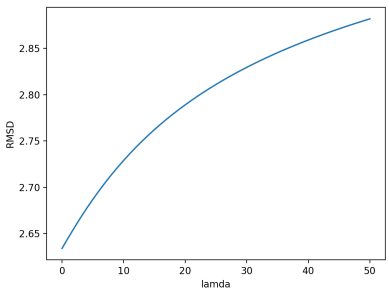
Part 1

a)



b) dimension 4 and dimension 6 stand out.
As $\lambda \to \infty$, $df(\lambda) \to 0$, dimension 6 is always larger than 0, which means it consistently gives positive prediction to the data and its magnitude is greater than 1,2,3,5,7 dims. It means it's more trustworthy than 1,4,3,5,7 features. Same as dim 4, which is always negative and its magnitude is larger than 1,2,3,5,7 dim. This means it's gives negative impact to the prediction and it's trustworthy and not very sensative to λ selection.

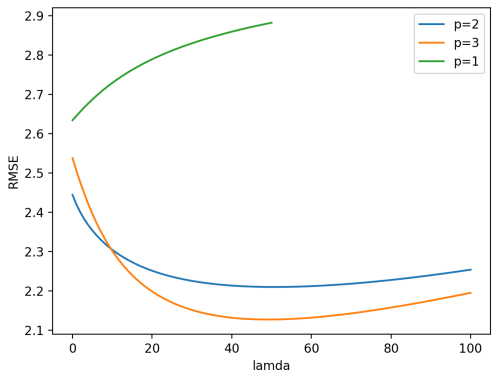
c)



As λ increases, the RMSE increases as well. This figure tells me we should set $\lambda = 0$ in this dataset, which means we should pick least square for this dataset.

part 2

d)



I would choose p=3 since it has the lowest RMSE at $\lambda = 40$. The ideal value of λ changes to $\lambda = 40$ for this problem.