# Homework 1

Sirui Tan
st2957

February 4, 2016

# 1 Problem 1

## 1.1 Part 1

### 1.1.1 (a)

$$L = \prod_{i=1}^{N} P(x_i|\pi)$$

### 1.1.2 (b)

Apply logarithm trick:

$$\log(L) = \sum_{i=1}^{N} \log(P(x_i|\pi)$$

Calculate gradient:

$$\Delta \log(L) = \sum_{i=1}^{N} \frac{P(x_i|\pi)'}{P(x_i|\pi)} = 0$$

Because

$$\frac{P(x|\pi)'}{P(x_i|\pi)} = \begin{cases} \dfrac{1}{\pi} & (x = 1) \\ \dfrac{-1}{1-\pi} & (x = 0) \end{cases}$$

Taken into gradient equation yields:

$$\frac{k}{\pi} = \frac{N-k}{1-\pi}$$

where $k$ represents the number of 1s in $\{x_1, \cdots x_N\}$. Therefore

$$\hat{\pi}_{ML} = \frac{k}{N}$$

#### 1.1.3 (c)

We can regard the observation set as taken from a sample space, where the definition of Bernoulli distribution indicates a portion $\pi$ of the samples are 1s. The i.i.d nature of the observations shows they are randomly taken from the sample space, so $\pi N$ of them are 1s is the most probable outcome.

### 1.2 Part 2

#### 1.2.1 (a)

$$L = \prod_{i=1}^{N} P(x_i|\lambda)$$

#### 1.2.2 (b)

Apply logarithm trick:

$$\log(L) = \sum_{i=1}^{N} \log(P(x_i|\lambda)$$

Calculate gradient:

$$\Delta \log(L) = \sum_{i=1}^{N} \frac{P(x_i|\lambda)'}{P(x_i|\lambda)} = 0$$

Because

$$\frac{P(x|\lambda)'}{P(x_i|\lambda)} = \frac{x-\lambda}{\lambda}$$

Taken into gradient equation yields:

$$\sum_{i=1}^{N} \frac{x_i - \lambda}{\lambda} = 0$$

Therefore

$$\hat{\lambda}_{ML} = \frac{\sum_{i=1}^{N} x_i}{N}$$

#### 1.2.3 (c)

From the definition of Poisson distribution, $x$ represents the number of events happen in a unit amount of time, and $\lambda$ represents the average number of event occurrences in that amount of time. Clearly the estimation is reasonable given the nature that samples are randomly taken.

# 2 Problem 2

## 2.1 (a)

Using Bayesian rule:

$$
\begin{aligned}
p(\lambda|x_1,\cdots,x_N) &\propto p(x_1,\cdots,x_N|\lambda)p(\lambda)\\
&\propto \prod_{i=1}^{N} p(x_i|\lambda)p(\lambda)\\
&\propto \lambda^{\sum x} e^{-\lambda}\lambda^{a-1}e^{-b\lambda}\\
&\propto \lambda^{(\sum x+a)-1}e^{-(b+N)\lambda}
\end{aligned}
$$

Clearly, the posterior of $\lambda$ is also a Gamma distribution.

$$\lambda \sim Gam(\sum x + a, N + b)$$

## 2.2 (b)

According to the definition of Gamma distribution:

$$
\begin{aligned}
E(\lambda) &= \frac{\sum x + a}{N + b}\\
var(\lambda) &= \frac{\sum x + a}{(N + b)^2}
\end{aligned}
$$

We can see here the mean and variance of the posterior distribution are both smoothed by factor $a$ in numerator and $b$ in denominator. So this relates to Part 2 of Problem 1 as being a normalized version of ML estimate.

# 3 Problem 3

## 3.1 Part 1

### 3.1.1 (a)

| Label | Value |
|---|---|
| intercept term | 23.5473136163 |
| number of cylinders | -0.53624543 |
| displacement | 1.27906133 |
| horsepower | -0.05922578 |
| weight | -6.19362742 |
| acceleration | 0.24448757 |
| model year | 2.83663356 |

Here, positive sign means input is proportional to output, negative sign means input is inverse proportional to output.

### 3.1.2 (b)

$$\mu_{MAE} = 2.67044377205$$
$$\sigma_{MAE} = 0.480117507216$$

## 3.2 Part 2

### 3.2.1 (a)

| $p$ | $\mu$ | $\sigma$ |
|---|---|---|
| 1 | 3.3875118672114795 | 0.68638502743433716 |
| 2 | 2.7864791778065361 | 0.66563803353712114 |
| 3 | 2.6891000472522428 | 0.6158636996172675 |
| 4 | 2.7235048725782884 | 0.61841079752286376 |

Clearly, $p = 3$ has the best performance.

### 3.2.2 (b)

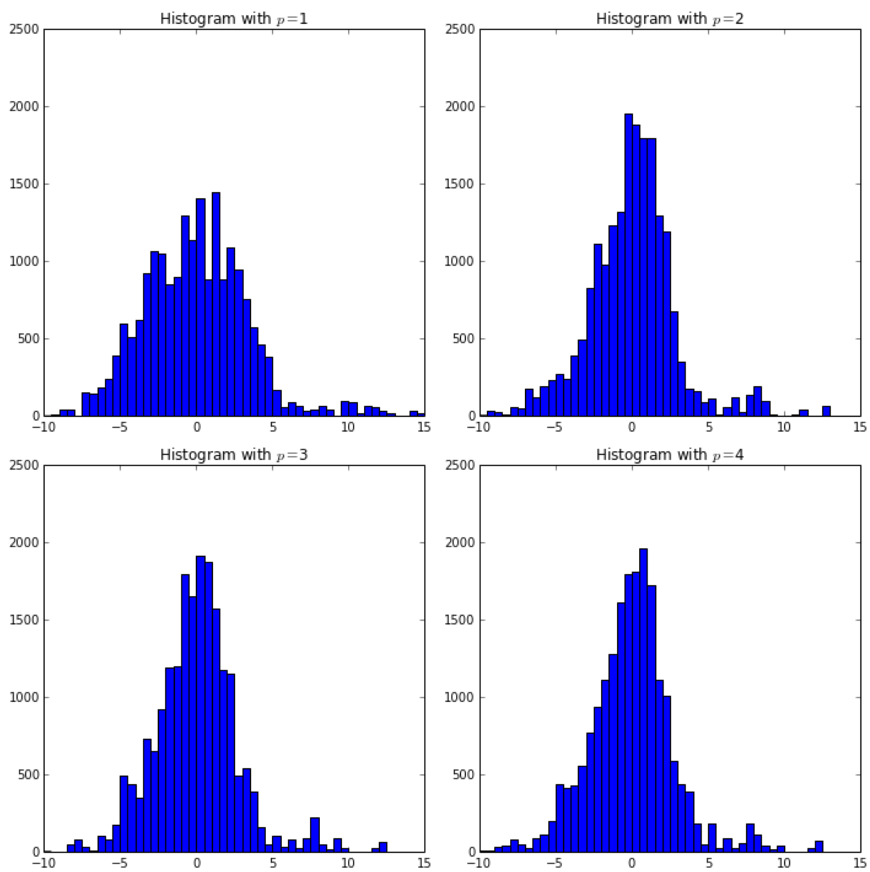


Figure 1: Histograms for each $p$ value

### 3.2.3 (c)

The maximum likelihood values of mean $\mu$ and variance $\sigma$ are calculated as

$$\mu = \frac{\sum_{i=1}^{20000} Err_i}{20000}$$
$$\sigma^2 = \frac{\sum_{i=1}^{20000} (Err_i - \mu)^2}{20000}$$

The log-likelihoods are

| $p$ value | log likelihood |
|---|---|
| 1 | -53292.1185271 |
| 2 | -49240.2680536 |
| 3 | -48466.7455654 |
| 4 | -49428.5843529 |

The results here agrees with that of part2(a) in that both show the optimal order is $p = 3$. The assumption that the log likelihood of errors is also a measure of the spread of the error range is satisfied. Meaning the larger the log likelihood is, the more concentrated the RMSE errors are, and the better the $p$ value is.