MLPR Assignment

I.D s0840844

March 12, 2013

0.1 Bayesian Analysis of Uniform distribution

Given $p(\theta) \sim Pa(b,K)$= $\begin{cases} \frac{Kb^k}{\theta^{K+1}} & \text{for } \theta \geq b \\ 0 & \text{otherwise} \end{cases}$

Now with the above Pareto prior, the joint distribution of θ and $D = \{x_1, x_2, ...x_N\}$ is:

$$p(D,\theta) = P(D|\theta)P(\theta) = \begin{cases} \frac{1}{\theta^N} \cdot \frac{Kb^K}{\theta^{K+1}} & \text{for } \theta \geq max(max(D), b) \\ 0 & \text{otherwise} \end{cases}$$

With the given definition of $P(D)$ and taking $m = max(D)$, the posterior $p(\theta|D)$ can now be defined as follows:

$$p(\theta|D) = \frac{P(\theta, D)}{P(D)} \quad = \begin{cases} \dfrac{\frac{1}{\theta^N} \cdot \frac{Kb^K}{\theta^{K+1}}}{\frac{K}{(N+K)b^N}} & \text{for } b \geq m \text{ and } b \leq \theta \\ \dfrac{\frac{1}{\theta^N} \cdot \frac{Kb^K}{\theta^{K+1}}}{\frac{Kb^K}{(N+K)m^{N+K}}} & \text{for } m > b \text{ and } m \leq \theta \end{cases}$$
$$0 \text{ otherwise} \tag{1}$$

This can be simplified and written as:

$$p(\theta|D) = \frac{P(\theta, D)}{P(D)} \quad = \begin{cases} \frac{(N+K)b^{N+K}}{\theta^{N+K+1}} & \text{for } b \geq m \text{ and } b \leq \theta \\ \frac{(N+K)m^{N+K}}{\theta^{N+K+1}} & \text{for } m > b \text{ and } m \leq \theta \end{cases} \tag{2}$$

Note : For b $\geq$ m , $p(\theta|D)$ =0 for $\theta < b$ and similarly for $m > b$ $p(\theta|D) = 0$ for $\theta < m$.
The posterior in fact has the same function functional form as a Pareto distribution :
We know the Pareto distribution, Pa is defined as : $Pa(\alpha, n) = \frac{\alpha n^\alpha}{\theta^{\alpha+1}}$. If we compare this with the above distribution, we can see $\alpha \equiv (N + K)$ and $n \equiv \{m, b\}$

0.2 The Tramcar problem

0.2.1 Part A

Assumptions made:

- trams are numbered sequentially as integers starting from 0 to some upper bound θ. Thus the likelihood function $p(x)$ can be defined as:

$$p(x) = f(x|\theta) = \begin{cases} \frac{1}{\theta} \text{ for } 0 \leq x \leq \theta \\ 0 \text{ otherwise} \end{cases}$$

- K=1 and b=1 and the D={100} . This implies m= 100

Since $m > b$ the posterior $p(\theta|D)$ is given as :

$$p(\theta|D) = \frac{(N+K)m^{N+K}}{\theta^{N+K+1}} = \frac{2.100^2}{\theta^3}$$

0.2.2 Part B

The posterior $p(\theta|D)$ as shown above is a Pareto distribution. Hence the mean and maximum posterior of $p(\theta|D)$ can derived using the properties of a Pareto distribution.

$$\mu(mean) = \frac{(N+K)m}{N+K-1} = \frac{2m}{1} = 200$$

$$MAP = mode = m = 100$$

0.2.3 Part C

The predictive density is given by $p(x|D) = \int_0^\infty p(x|\theta)p(\theta|D)d\theta$.
As already stated in part A , $p(\theta|D) = 0$ for $\theta < m$ thus

$$\begin{aligned} p(x|D) &= \int_m^\infty p(x|\theta)p(\theta|D)d\theta \\ p(x|D) &= \int_{100}^\infty \frac{1}{\theta}.\frac{2.100^2}{\theta^3}d\theta \\ &= [\frac{-2.100^2}{3.\theta^3}]_{100}^\infty \\ &= \frac{2}{3.100} \\ &= \frac{1}{150} \end{aligned} \tag{3}$$

0.2.4 Part D

From Part C, it is clear that the predictive distribution $p(x|D)$ is a uniform distribution $U(x, \theta)$ where $\theta = 150$:

$$p(x|D) = \begin{cases} \frac{1}{150} \text{ for } 0 \leq x \leq 150 \\ 0 \text{ otherwise} \end{cases}$$

Therefore for a new data point $\mathbf{x}$, the prediction is :

$$p(\mathbf{x}|D) = \frac{1}{150} I(\mathbf{x} \in [0, 150]) = \frac{1}{150} I(\mathbf{x} \leq 150)$$

For observations whose value lie outside 150, the probability of observing them given the dataset D is 0. Thus $p(50|D) = \frac{1}{150}$ and $p(500|D) = 0$.

0.2.5 Part E

As K $\longrightarrow$ 0
$\lim_{K\to 0} p(\theta) = 0$. Thus the limit of posterior $p(\theta|D)$ when K tends to 0 :

$$\lim_{K\to 0} p(\theta|D) = \frac{P(\theta, D)}{P(D)} = \begin{cases} \frac{(N)b^N}{\theta^{N+1}} \text{ for } b \geq m \text{ and } b \leq \theta \\ \frac{(N)m^N}{\theta^{N+1}} \text{ for } m > b \text{ and } m \leq \theta \end{cases}$$

Observations:

- As $K \to 0$, the posterior $p(\theta|D)$ at the limiting value of K is still a Pareto distribution. Sending K to 0 tends to change only the shape parameter of the original posterior distribution which is illustrated as follows:

  We know $Pa(\kappa, n) = \frac{\kappa n^\alpha}{\theta^{\kappa+1}}$.

  By comparing the above distribution with our original posterior $p(\theta|D)$, we concluded that $\kappa \equiv (N + K)$ but as K$\to$ 0 $\kappa \equiv N$

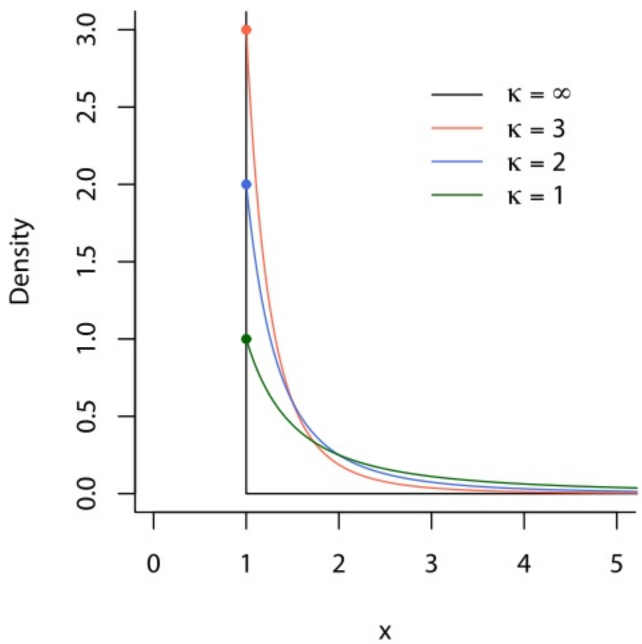
  By observing the graph below, we can see that decreasing the value of the shape parameter κ makes the distribution more heavily tailed. Thus, rare instances will have a greater probability mass assigned to them than before. Having the posterior $p(\theta|D)$ more heavily tailed improves the predictive density that we computed in Part C.

  $$\lim_{k\to 0} p(x|D) = \int_{100}^{\infty} \frac{1}{\theta} . \frac{1.100^1}{\theta^2} d\theta = \frac{1}{200}$$

  Therefore for a new data point $\mathbf{x}$, the prediction is :

  $$p(\mathbf{x}|D) = \frac{1}{200} I(\mathbf{x} \in [0, 200]) = \frac{1}{200} I(\mathbf{x} \leq 200)$$

  Although the individual probability of observing each individual tram within the range [0-150] is less than before, by assigning probability masses on trams between 150 and 200, the predictor improves prediction by increasing the set of trams that it believes that we might see next. In other words the predictor moves towards the true uniform distribution.



Density
x
κ = ∞
κ = 3
κ = 2
κ = 1