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 $\kappa \equiv N$

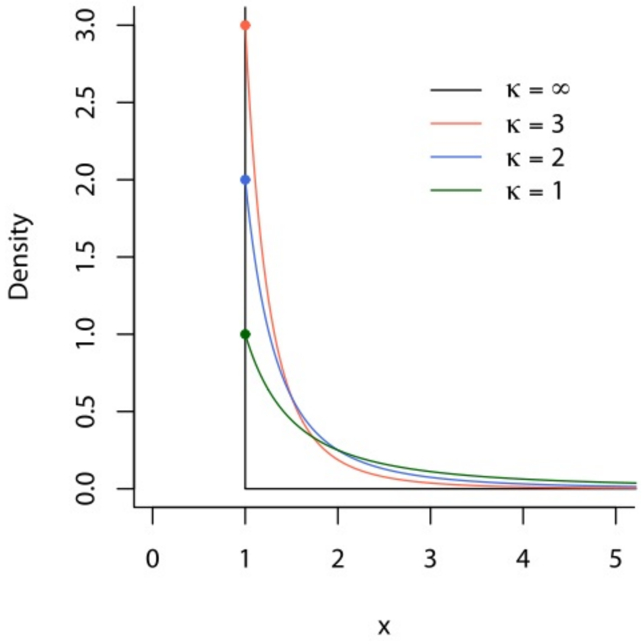
  By observing the graph below (where x denote θ), we can see that decreasing the value of the shape parameter κ makes the distribution more heavily tailed. Thus, rare instances will have a greater probability mass assigned to them than before. Having the posterior $p(\theta|D)$ more heavily tailed improves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



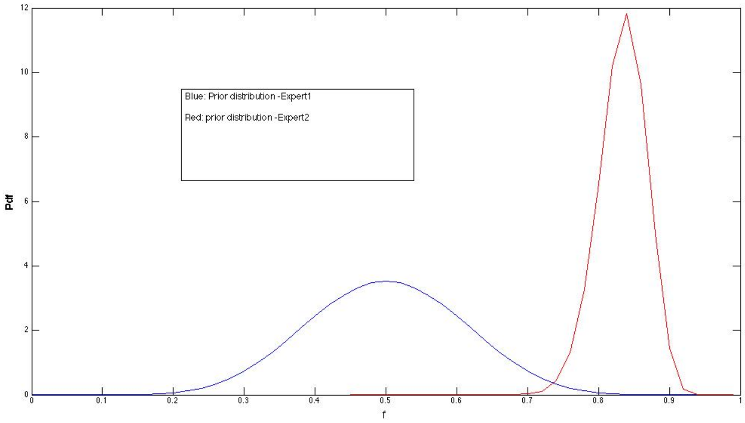
0.3 Bayesian Classification

Given :
Initially expert1's knowledge regarding the probability of a food delivery being Soylent Red rather than Soylent Yellow is given by a beta distribution with parameters $(n_0, n_1) = (10, 10)$ while expert2's initial knowledge is best described by the beta distribution with parameters $(n_0, n_1) = (100, 20)$.
Hence expert 1's prior belief is captured by $p(f) = f^{(10)-1}(1-f)^{(10)-1}$ where f is probability of seeing Soylent Red. and expert 2's corresponding prior belief is given by $p(f) = f^{(100)-1}(1-f)^{(20)-1}$.
The prior distributions of both experts is shown in figure1.
Observation

- Since the true value of f is $\frac{3}{4}$, it can be seen that expert 2 initially has a better opinion.

  Reason: Without seeing any data, if both experts were asked to predict the probability Soylent Red, then on average both experts will chose a value of f that is equal to the mean of the prior distribution. The mean of expert 1's prior distribution is $\frac{1}{2}$ whereas for expert2 the mean of the beta distribution is $\frac{100}{120}$. The expected value of the prior distribution for expert2 is closer to the true value. Hence, judging from these values taken by the means it can be observed that expert 1 believes that on average he will see equal number of Soylent Reds and Soylent yellows whereas Expert2 believes on average he will see more Soylent reds than Soylent yellows.

Figure 1: Prior beliefs of Expert1 and Expert2.



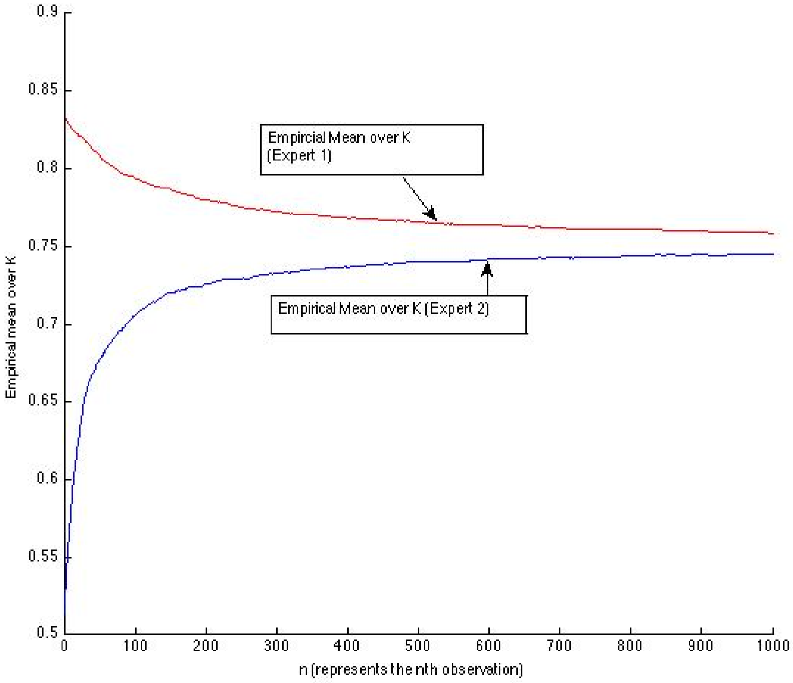
0.3.1 Computing the posterior

With incoming data, the posterior distribution of both experts is given by:

$$\begin{aligned} p(f|\mathbf{X}) &\propto p(\mathbf{X}|f)p(f) \\ &= f^k(1-f)^{N-k}f^{n_0-1}(1-f)^{n_1-1} \\ &= f^{(k+n_0)-1}(1-f)^{(N+n_1-k)-1} \end{aligned} \tag{4}$$

here k represents the number of food items seen that correspond to Soylent red and N corresponds to the total number of observations that has been made so far..
For Expert 1, the corresponding posterior distribution is given by:
$p(f|X) = f^{(k+10)-1}(1-f)^{(N+10-k)-1}$
And similarly, for expert 2 the corresponding posterior distribution is :
$p(f|X) = f^{(k+100)-1}(1-f)^{(N+20-k)-1}$
Note: The mean and mode of the posterior distributions of both experts change as more observations made. The mean of the posterior distribution for expert 1 is $\mu_1 = \frac{K+10}{N+20}$ and similarly the mean of the posterior distribution of expert 2 is $\mu_1 = \frac{K+100}{N+120}$. The terms K and N change with more data arrives. Hence, we can conclude that the expectation of seeing a Soylent red rather than Soylent yellow change as they continue seeing new data. This is illustrated by the graph in figure 2.
Observation: The mean of the posterior distribution $p(f|X)$ for both experts converge toward the true value of f as the number of observations increase. This is because the two parameters k and N tends to dominate the numerator and denominator in the expression used to compute the mean $\mu = \frac{k+n_0}{N+n_1}$ as the number of samples increase.

Figure 2: Empirical posterior mean μ over k against N



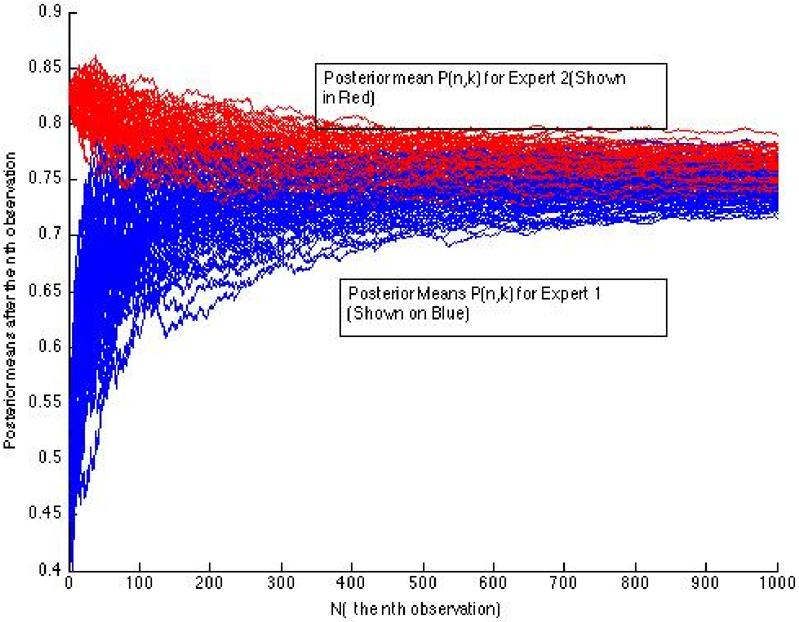
0.3.2 Nature of Convergence

Observation: On careful observation of the plots given by figure 3(where the posterior mean of each expert at the kth simulation, computed after the nth observation, is presented by the y axis), it can be deduced that the posterior distribution of expert 1 is more responsive to incoming data than that of expert 2's posterior distribution.
Reason : The mean and mode of a beta distribution is given by:

$$mu = \frac{k + n_0}{N + n_1} \text{ and } mode = \frac{k + n_0 - 1}{N + n_1 - 2}$$

The parameter $n_0 >> n_1$ in expert 2's posterior distribution and this in fact has a substantial impact on the beliefs of the second expert. Since $n_0 >> n_1$, a large of number of observations have to be made to see substantial changes in the value of the mode and mean of the expert 2's posterior distribution. But this is not the case for expert 1 where $n_0 = n_1$, Hence the sharp initial slopes of the blue plots after the initial observations clearly show that expert 1's posterior distribution is more sensitive to the data. Further evidence to support this claim can found by observing that the blue plots saturate towards the true value earlier than the red plots.

Figure 3: The posterior mean p_{exp} for each k against N



0.3.3 Uncertainty in the long run

Observation : From the observation of the plots shown by figure 4 and figure 5, it is clear that the degree of uncertainty regarding the value of f decreases in the long run. This can seen from the decrease in the size of the standard deviation and variation as N increase. The posterior distribution of both experts thus become narrower with increasing N and their means move slowly toward $f = 0.75$ with the increasing number of observations. Intuitively, both experts become more and more sure about the true probability distribution as they encounter more data.

Figure 4: The mean ($\pm$) standard deviation of the posterior at k=1 against N

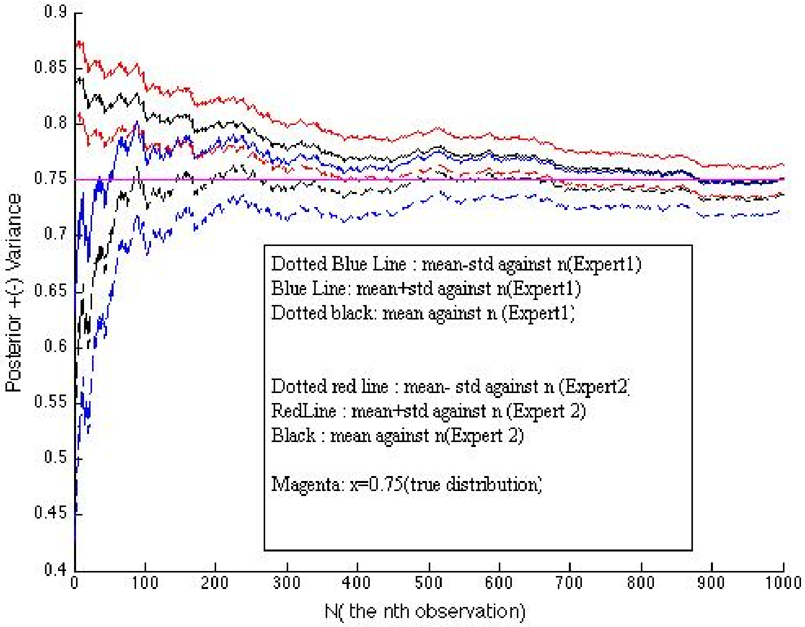
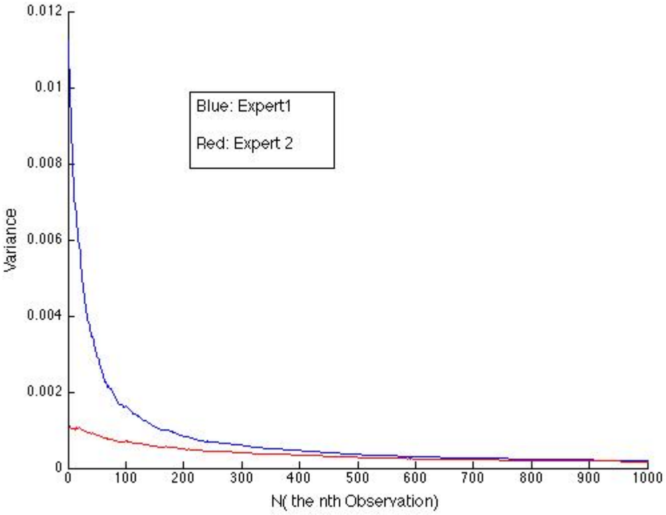


Figure 5: The variance of the posterior at k=1 against N



0.4 Bayesian regression

The models considered in this last section are linear regression models. The main property of such models is that the prediction $f(x;w)$ is a linear function of weights

$$f(x,w) = \sum_{j=1}^{j=M} w_j \phi_j(x) + w_0$$

where $\phi_j(x)$ are basis functions
The above expression can be written in matrix form by introducing a dummy basis function $\phi_0(x) = 1$ to account the parameter w_0 :

$$f(x,w) = \sum_{j=0}^{j=M} w_j \phi_j(x) = \mathbf{w}^T \Phi$$

where $\Phi = (\phi_0, \phi_1 ... \phi_M)$ and $\mathbf{w} = (w_0, w_1 .... w_M)$(M denotes the number of basis functions used)
We make an assumption that the target variable y can be explained by the linear model $f(x.w)$ with additive gaussian noise so $y = f(x;w) + \epsilon$. Hence, given our training set of inputs $X = x_1, ..., x_N$ with corresponding target values $y1, ..., y_N$ and making the assumption that the points have been drawn independently, the log- likelihood function can be presented as :

$$ln(p(y|X,w,\beta)) = \sum_{n=1}^{N} ln(N(y_n|w^T \Phi(x_n), \beta^{-1}) = constant - E_D(w)$$

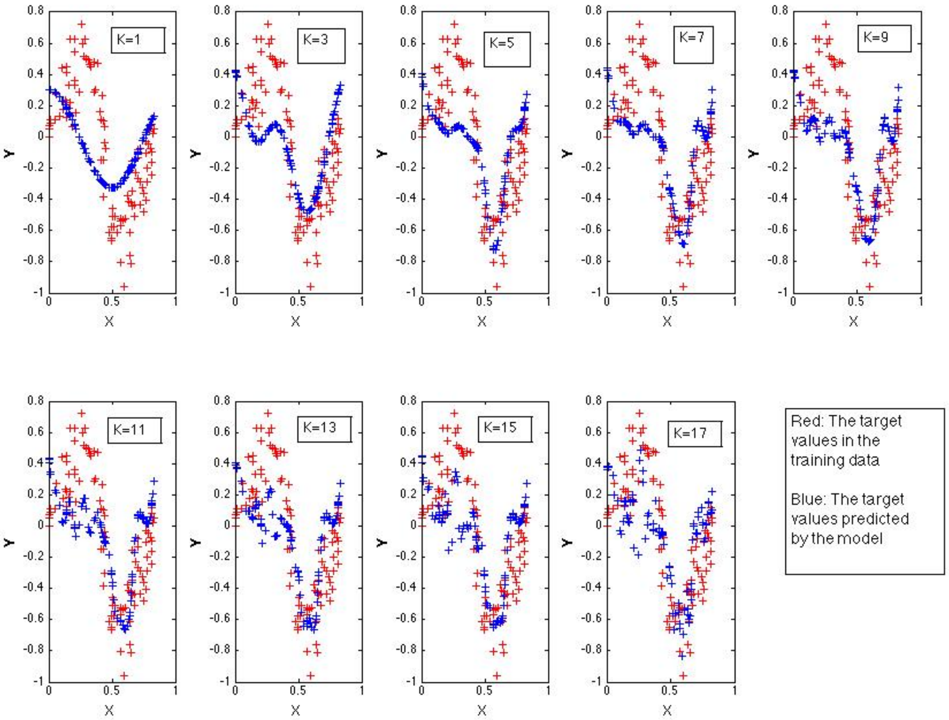
where $E_D(w) = \frac{\beta}{2} \|\mathbf{y} - \Phi \mathbf{w}\|^2$. Thus, maximising the log-likelihood is hence equivalent to minimising $E_D(\mathbf{w})$ with respect to w . A closed-form solution exists for this function

$$w = (\Phi^T \Phi)^{-1} \Phi^T Y$$

0.4.1 Part A

The object of Part A is to compare the performance of the prediction $f(x;w)$ on the training data given different linear models. Each model M_i is distinguished by the number of basis functions the model employs i.e M. (All model employ fourier basis functions). From the observation of figure 6, it can be seen that increasing k (i.e M in our case) results in the use of linear models that employ more basis functions resulting into better fits of the data. Initially, this proves to be a good heuristic step since since increasing the model complexity (i.e the number of free parameters) results into flexible models that not only give better fits to the training data but also result in better representations of the true function. However, the training data only represents a more portion of the true data space, hence increasing M will inevitably lead to models that overfit the training data. This normally results due to the fact as M increases, the weights become more finely tuned to the training data by taking large positive and negative values.

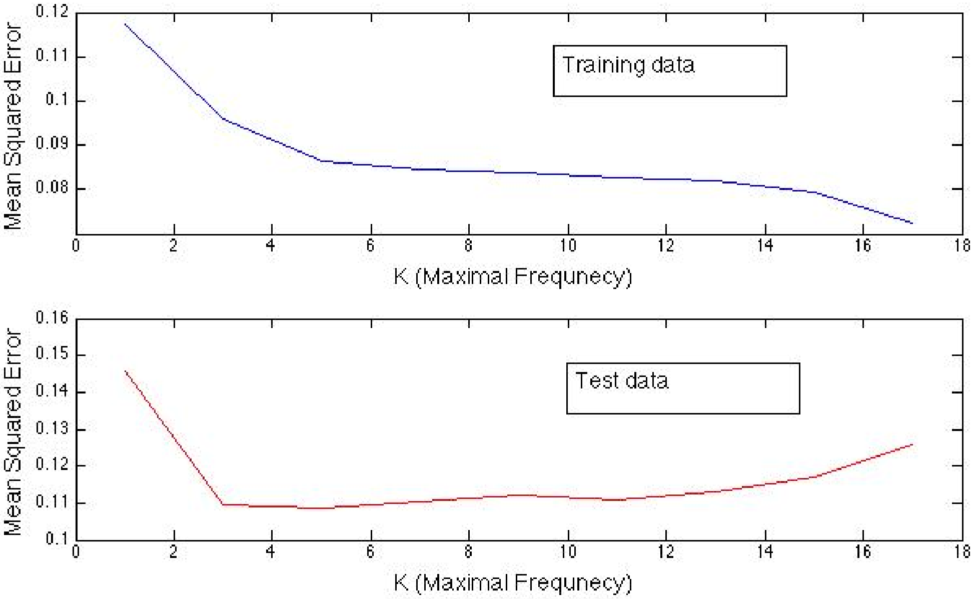
Figure 6: Functional fits together with the training data against k



0.4.2 Part B

Figure 7 shows the typical behaviour of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In that case the predictions $f(x : w)$ will have large variance thus explaining the growth of the test error in the latter stages. In contrast, if the model is not complex enough, it will underfit and may have large bias, again resulting in poor generalization. This is explained by the high test error observed for very very simple models.

Figure 7: The Mean squared error against maximum frequency



0.4.3 Part C

As we have seen in Part A and B, using maximum likelihood leads to the choice of parameters **w** that overfits the data. Through the introduction of the Gaussian prior on the weight, we incorporate a regularisation term that controls the degree of over fitting. The motivation behind this is as follows:
We make an assumption that the target variable y can be explained by the linear model $f(x.w)$ with additive gaussian noise so $y = f(x;w) + \epsilon$ The evidence therefore $P(D|M_i)$ is given by

$$p(D|M_i) = \int p(D, w, M_i)dw = \int P(D|w, \beta)p(w|\lambda, M_i)dw$$

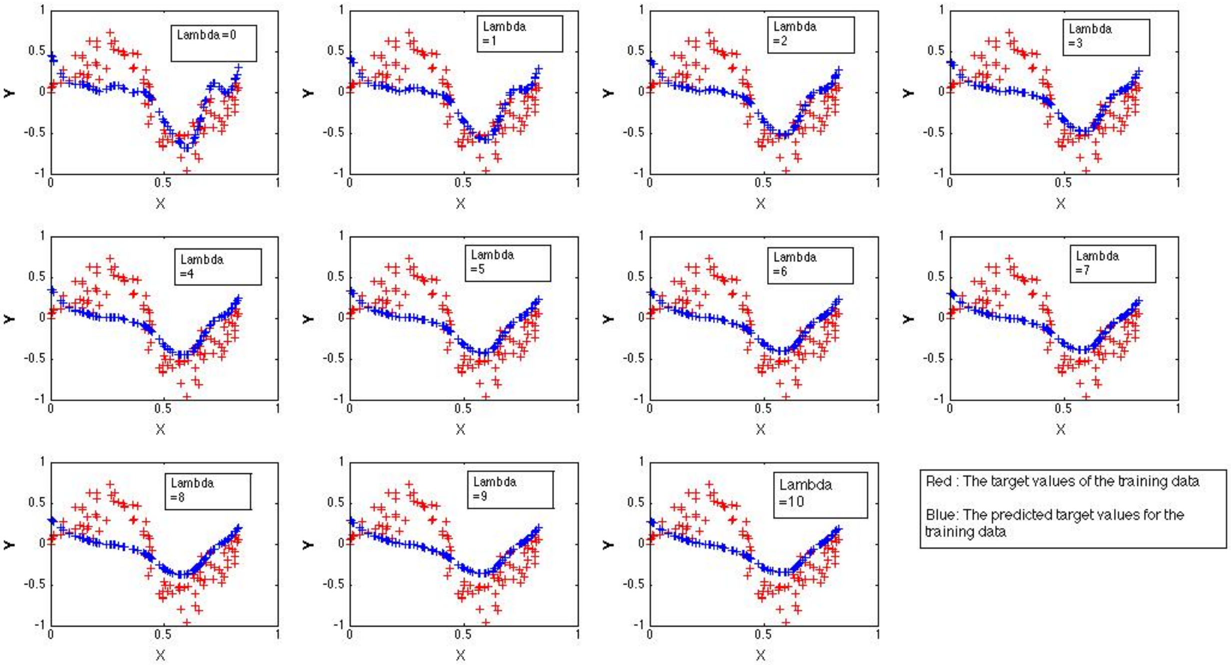
. (Here D represents the observed target values for which the corresponding inputs are known). The marginal likelihood can thus be viewed as the probability of generating the data set D from a model whose parameters are sampled at random from the prior. By keeping the hyper parameters fixed, we compute an approximation of $p(D|M_i)$ by using **w** for which $P(D|w, \beta)p(w|\lambda, M_i)$ is maximum. The log-likelihood of this approximate $P(D|M_i)$ is:

$$lnP(D|M_i) = constant - E(w)$$

where $E(w) = \frac{\beta}{2}\|\mathbf{y} - \Phi\mathbf{w}\|^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$ Similar to problem formulation in Part A, the problem of maximising the log-likelihood is now equivalent to minimising E(**w**).
Note : λ controls the relative importance of the data dependent error and the regularisation term. Increasing the value of this parameter encourages weight values to decay to 0 thus decreasing the

complexity of the model. This counteracts with the large values that set for **w** when we increase the number of basis functions. From the observation of figure 8, it can be seen that as λ increases, the complexity of the model decreases. Hence, the the number of bends(local minima) in the predictive function slowly starts to decrease with increasing .

Figure 8: Functional fits using Gaussian prior together with training data against k



0.4.4 Part D

As mentioned earlier, the log of the marginal likelihood of each model can be represented as :

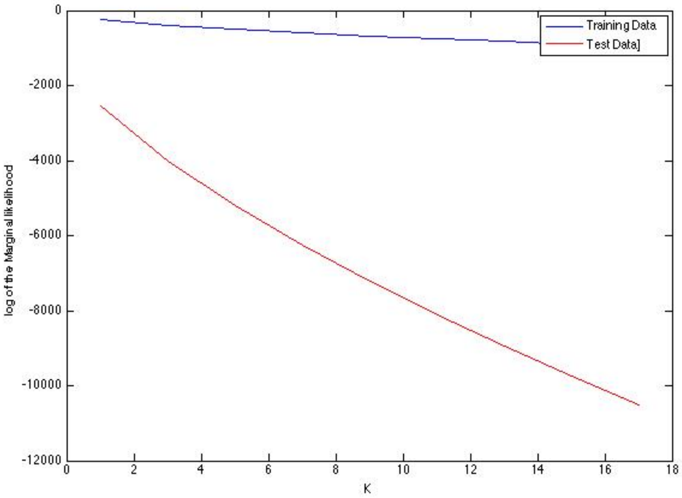
$$ln(p(D|M_i)) = ln(\int p(D, w, M_i)dw) = ln(\int P(D|w, \beta)p(w|\lambda, M_i)dw)$$

Figure 9 shows the plot of the log of the evidence against the models M_i (As mentioned before, each model M_i is only distinguished by its value k) From the plots, it can be seen that the log of the evidence decreases with increasing k for both the training and test data. The decrement of the log-likelihood is more pronounced on the test data than on the training data. When choosing models using bayesian statistics, the model M_i is chosen which has the highest $P(M_i|D)$ where

$$P(M_i|D) \propto P(D|M_i)P(M_i)$$

Since we are equally uncertain about each model, assuming $P(M_i)$ to be a uniform seem reasonable in our problem. Thus we choose the model that has the highest $P(D|M_i)$. Therefore from the plots, it can be observed that simple models on general gives better representation of the data than complex models.

Figure 9: log of the evidence against k


Training Data
Test Data]
log of the Marginal likelihood
K