CS215: Data Analysis and Interpretation

Assignment 5

BY:	*Roll Number*
Harshit Gupta	190050048
Pradipta Parag Bora	190050089

Contents

Question 1

Answer:

1. We have implemented the algorithm for generating samples in the code. It is well documented, kindly refer to it for implementation details. Here we will derive the expressions for both the MAP estimates and the MLE estimates which we are using in the code.

   Let the vector of N samples be called as X. Then the posterior is:

$$P(\mu|X) = \frac{P(X|\mu)P(\mu)}{\int_\mu P(X|\mu)P(\mu)d\mu}$$

   The likelihood $P(X|\mu)$ is:

$$P(X|\mu) = \prod_{i=1}^{N} P(X_i|\mu) = \prod_{i=1}^{N} G(X_i, \mu, \sigma^2)$$

   where $G(X, \mu, \sigma^2)$ is the pdf of the normal distribution with mean μ and variance σ^2. As done in the slides the product can be rewritten in the form (after some algebra):

$$P(X|\mu) = \prod_{i=1}^{N} G(X_i, \mu, \sigma^2) \propto G(\mu, \bar{X} = \frac{\sum_i X_i}{N}, \sigma^2/N)$$

   For the MLE estimate we just need to maximise the likelihood which for a gaussian happens at the mean. Thus:

$$\mu^{MLE} = \bar{X} = \frac{\sum_i X_i}{N}$$

   For Bayesian analysis the posterior now becomes:

$$P(\mu|X) = \frac{G(\mu, \bar{X} = \frac{\sum_i X_i}{N}, \sigma^2/N)P(\mu)}{\int_\mu G(\mu, \bar{X} = \frac{\sum_i X_i}{N}, \sigma^2/N)P(\mu)d\mu}$$

   For MAP the denominator does not matter (as it is not a function of μ). Thus:

$$\left.\frac{\partial G(\mu, \bar{X}, \sigma^2/N)P(\mu)}{\partial \mu}\right|_{\mu=\mu^{MAP}} = 0$$

   Now for the two priors given:

   (a)

$$P(\mu) = G(\mu, \mu_0, \sigma_0^2) = G(\mu, 10.5, 1)$$

   In this case the numerator will be:

$$G(\mu, \bar{X}, \sigma^2/N)P(\mu) * G(\mu, \mu_0, \sigma_0^2) \propto G(\mu, a, b)$$

   where

$$a = \frac{\bar{x}\sigma_0^2 + \mu_0\sigma^2/N}{\sigma_0^2 + \sigma^2/N}$$

   The MAP estimate will be the mean of the product gaussian which is a. Thus:

$$\mu^{MAP1} = \frac{\bar{x}\sigma_0^2 + \mu_0\sigma^2/N}{\sigma_0^2 + \sigma^2/N}$$

   with $\mu_0 = 10.5$ and $\sigma_0 = 1$

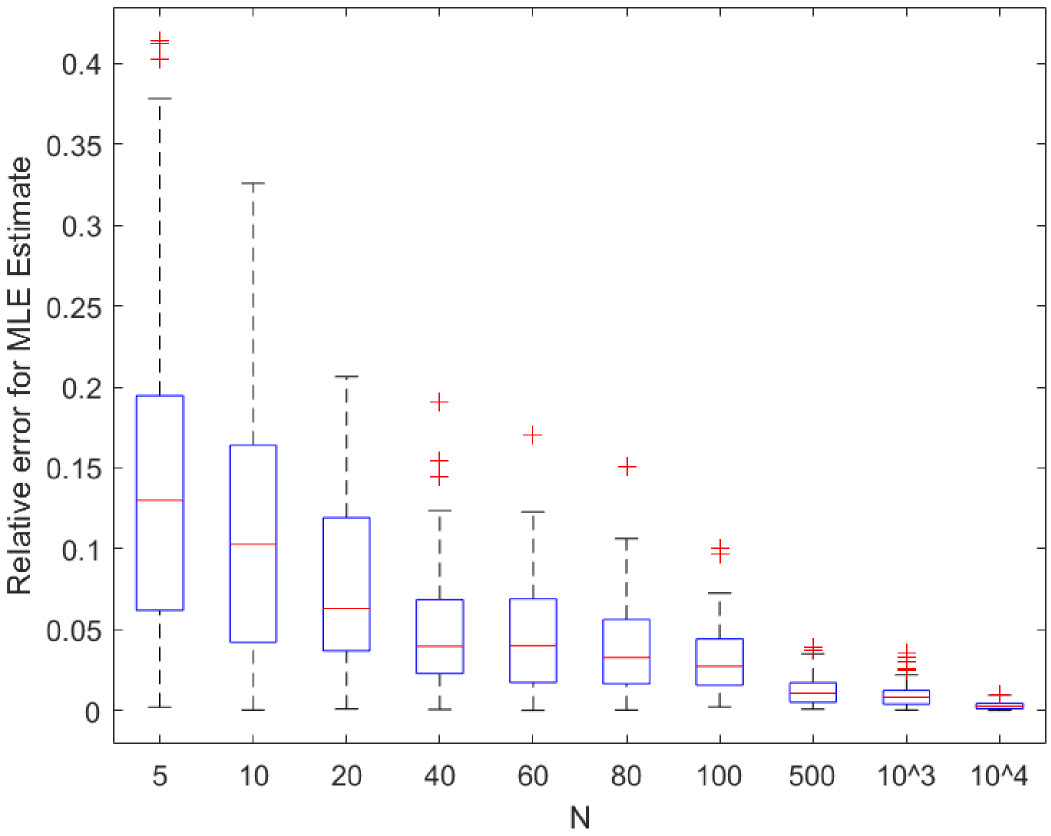
(b) In the second prior we have when $9.5 \leq \mu \leq 11.5$,

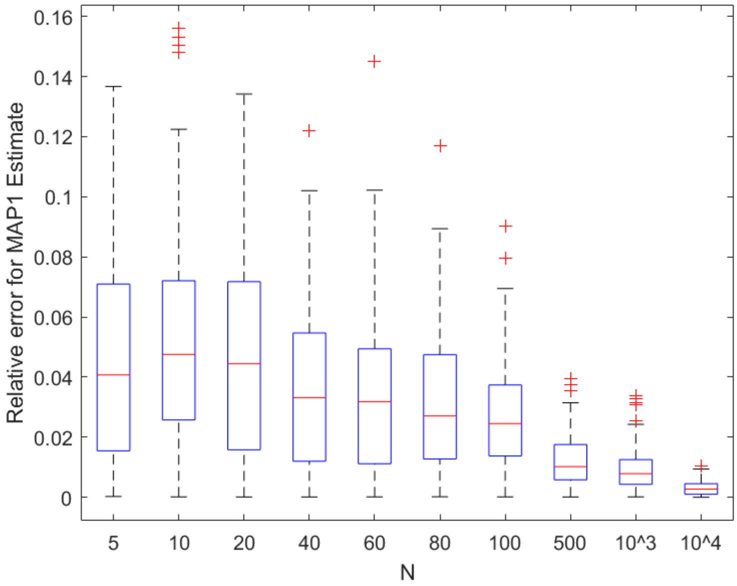
$$P(\mu) = 0.5$$

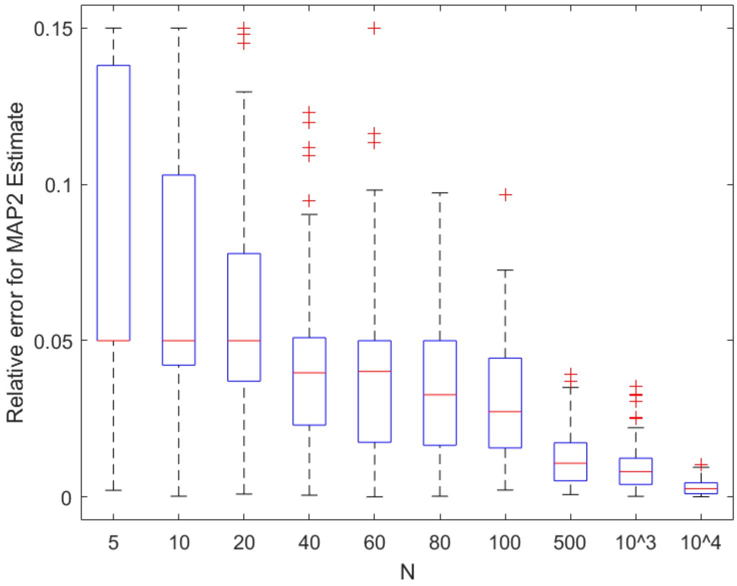
and otherwise the density is 0. Thus this case is almost the same as the MLE estimate since we are almost maximising the likelihood. However if the sample mean lies outside the range $9.5 \leq \bar{x} \leq 11.5$ then its not sensible to use this value since then at the sample mean the likelihood is 0. In this case if the sample mean $\bar{x} < 9.5$ then we will have 9.5 as the MAP estimate since it will give the highest value as the function in the numerator is of the form $-(\mu - \bar{x})^2$ and so the closest valid value to $\bar{x}$ would be 9.5 which will give the highest likelihood. Similarly if $\bar{x} > 11.5$ then we will have 11.5 as the MAP estimate. Thus:

$$\mu^{MAP2} = \begin{cases} \bar{x} & 9.5 \leq \bar{x} \leq 11.5 \\ 9.5 & \bar{x} < 9.5 \\ 11.5 & 11.5 < \bar{x} \end{cases}$$

2. The graphs are as follows:







3. We clearly see that MAP estimates are clearly more accurate than the MLE estimate especially at lower sample sizes. This is because we are leveraging the prior information in deciding the value of the parameter which helps reduce the variability and bias especially at lower sample values. For higher values of sample size all three estimators reach the true estimate.

   It seems that the gaussian prior performs better even on sample sizes although0 it does have some outliers with high (> 0.1 relative error) But mostly it performs the best with its median showing the least errors. The uniform prior also works well (but not as well as the gaussian) and this could be attributed to the fact that all numbers in $[9.5, 11.5]$ are equivalent for the prior.

   (a) The error decreases in all three estimators as N increases and for very large N the error is negligible.

(b) As explained above I would prefer μ^{MAP1} because this estimator performs the best even at low sample sizes. This is because the prior which is a gaussian assigns high density to values close to the true mean and so helps in getting an estimate with low error. The uniform prior works well compared to the MLE estimator but since it treats all values in $[9.5, 11.5]$ equivalently it has less information and is less power than the gaussian prior. So we should use the gaussian prior.

Instructions for Code: Run `q1.m` for this problem. This will display the three plots that we have displayed above. The plots have also been saved in the results folder of this question for reference.

Question 2

Answer:

1. Let us first derive the expression for the transformed distribution. Let the initial distribution be $p(x) = 1$ for $0 \leq x \leq 1$. We have the following transformation:

$$y = f(x) = (-1/\lambda)\log(x)$$

This gives us:

$$x = f^{-1}(y) = \exp(-\lambda y)$$

for $y \in [0, \infty)$ By the formula for transformation of random variables, we get the new distribution $q(y)$ as:

$$q(y) = p(f^{-1}(y))|\frac{df^{-1}(y)}{dy}| = \lambda \exp(-\lambda y)$$

When we sample N data points $X_1, X_2 \cdots X_N$ from this distribution the likelihood is:

$$P(X_1, X_2 \cdots X_N|\lambda) = \lambda^N \exp(-\lambda \sum_i X_i)$$

We define $\sum_i X_i$ as s to simplify the expressions. Differentiating the negative log likelihood to get the MLE:

$$\left.\frac{\partial NLL}{\partial \lambda}\right|_{\lambda=\lambda^{MLE}} = -N/\lambda^{MLE}) + S = 0$$
$$\implies \lambda^{MLE} = \frac{N}{S} = \frac{N}{\sum_{i=1}^{N} X_i}$$

2. Let us now derive the posterior mean. We have a gamma PDF as the prior. Let X denote the current sample. The expression for the posterior is:

$$P(\lambda|X) = \frac{P(X|\lambda)P(\lambda)}{\int_\lambda P(X|\lambda)P(\lambda)d\lambda} = \frac{\lambda^N \exp(-\lambda S)\beta^\alpha/\Gamma(\alpha)\lambda^{\alpha-1}\exp(-\lambda\beta)}{\int_\lambda \lambda^N \exp(-\lambda S)\beta^\alpha/\Gamma(\alpha)\lambda^{\alpha-1}\exp(-\lambda\beta)d\lambda}$$
$$= \frac{\lambda^{N+\alpha-1}\exp(-\lambda(S+\beta))}{\int_0^\infty \lambda^{N+\alpha-1}\exp(-\lambda(S+\beta))d\lambda}$$

The Posterior mean therefore will be:

$$\lambda^{Mean} = \int_0^\infty \lambda P(\lambda|X)d\lambda = \frac{\int_0^\infty \lambda^{N+\alpha}\exp(-\lambda(S+\beta))d\lambda}{\int_0^\infty \lambda^{N+\alpha-1}\exp(-\lambda(S+\beta))d\lambda}$$

We will use the following result:

$$I = \int_0^\infty \lambda^a \exp(-b\lambda)d\lambda = \frac{\Gamma(a+1)}{b^{a+1}}$$

This can be proven by substitution. Substitute $u = b\lambda$. Then the integral becomes:

$$I = \int_0^\infty (u/b)^a \exp(-u)du/b = \frac{1}{b^{a+1}}\int_0^\infty u^a \exp(-u)du = \frac{\Gamma(a+1)}{b^{a+1}}$$

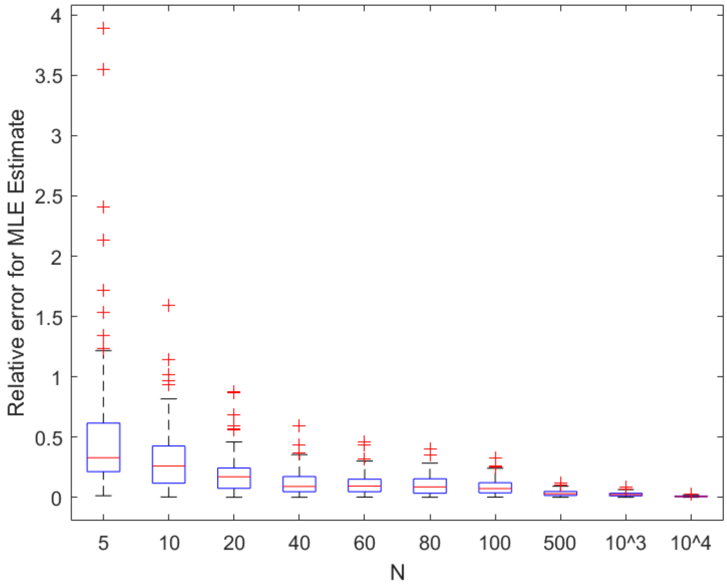
Using this we get:

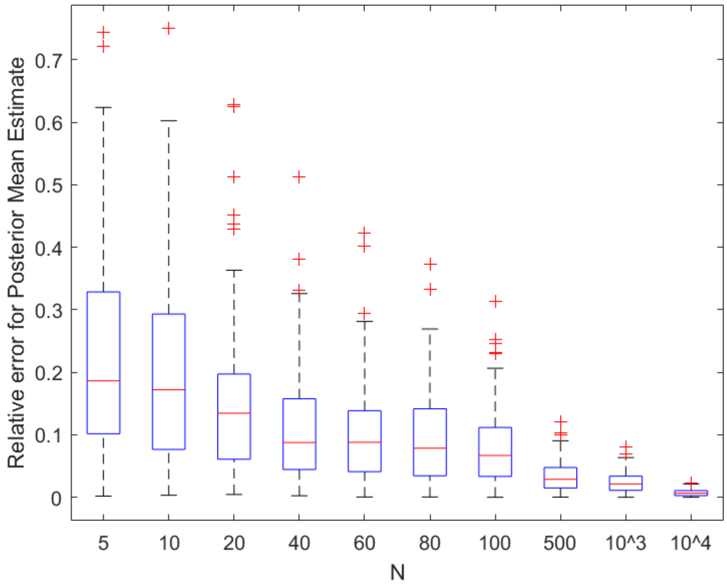
$$\lambda^{Mean} = \frac{\int_0^\infty \lambda^{N+\alpha} \exp(-\lambda(S+\beta))d\lambda}{\int_0^\infty \lambda^{N+\alpha-1} \exp(-\lambda(S+\beta))d\lambda} = \frac{\Gamma(N+\alpha+1)/(S+\beta)^{N+\alpha+1}}{\Gamma(N+\alpha)/(S+\beta)^{N+\alpha}}$$

$$= \frac{\Gamma(N+\alpha+1)}{(S+\beta)\Gamma(N+\alpha)} = \frac{N+\alpha}{S+\beta}$$

We are using this formula in the code to estimate λ.

3. The graph of the relative error for both the estimates are:





4. We can see that once again the Posterior Mean estimator is working far far better than the MLE estimator especially at lower number of samples. This is again because the posterior mean estimator has the terms of α and β in the numerator and denominator respectively which provide an initial baseline (which we know from the prior) and hence for low number of samples, is more accurate and variable.

   (a) Again the error decreases in both estimators as N increases and for very large N the error is negligible. Thus the estimators are converging to the true value.

   (b) We should prefer the bayesian posterior mean estimator as it shows far lower relative errors than the MLE estimator at lower number of sample. Since in practical cases we do not have access to a large amount of samples, using the bayesian analysis method with a proper prior is more helpful especially as it allows us to leverage our previously learnt prior models. This is seen from the graphs where the MLE estimator in some cases has relative errors as high as 4% with lesser number of samples whereas the Posterior mean estimator is consistently better performing with relative error $< 1\%$ even on smaller samples.

   Instructions for Code: Run `q2.m` for this problem. This will display the two plots that we have displayed above. The plots have also been saved in the results folder of this question for reference.