CS215 Assignment 3

210050044 Dhananjay Raman
210010076 Shantanu Welling

October 2022

Contents

Problem 1

1. Plot of the relative errors for each estimate:

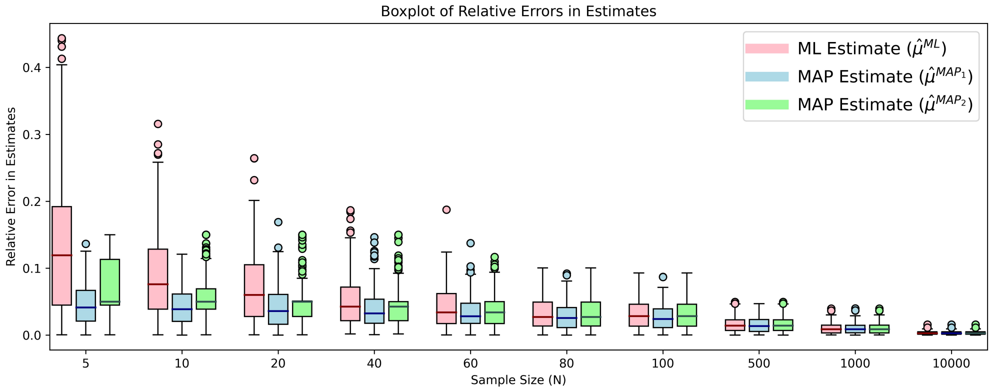

Figure 1: Relative errors for ML estimate, and the two MAP estimates

The maximum likelihood estimate $\hat{\mu}^{ML}$ would be sample mean for a Gaussian likelihood function. For a Gaussian prior:
Prior: $P(\mu) = G(\mu; \mu_{prior} = 10.5, \sigma^2_{prior} = 1)$
Likelihood: $P(data|\mu) = \prod_i G(x_i; \mu, \sigma^2_{true} = 16)$ $\quad$ ($\{x_i\}_{i=1}^N$ is the sample data)
$\prod_i G(x_i; \mu, \sigma^2_{true} = 16) = \prod_i G(\mu; x_i, \sigma^2_{true} = 16)$

Thus, likelihood is proportional to $G(\mu;\ \sum_i^N x_i/N,\ \sigma^2_{true}/N)$.
Product of 2 Gaussian PDFs is $G(z;\mu_1,\sigma_1^2)G(z;\mu_2,\sigma_2^2) = G(z;\mu_3,\sigma_3^2)$, where

$$\mu_3 = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2},\ \sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

So, Posterior PDF is proportional to Likelihood × Prior, which is Gaussian (as it is the product of 2 Gaussian PDFs)
Therefore, MAP estimate $\hat{\mu}^{MAP_1}$ is the posterior PDF's location parameter, i.e.

$$\hat{\mu}^{MAP_1} = \frac{\bar{x}\sigma^2_{prior} + \mu_{prior}\sigma^2_{true}/N}{\sigma^2_{prior} + \sigma^2_{true}/N}\ ; \qquad (\bar{x} = \sum_i^N x_i/N)$$

For a uniform prior:
If the mean of likelihood Gaussian PDF (sample mean, which is also the mode as the PDF is Gaussian) lies in the range of uniform prior, then it itself is the MAP estimate as in that range, the likelihood would only be scaled by a constant factor, the mode (maxima) of the PDF **would not change**.
If the sample mean of likelihood Gaussian PDF is greater than the right limit of the uniform prior, then the posterior PDF (which is the likelihood times some constant probability of the uniform prior on that range) is monotonically increasing on the range of uniform prior. This is because the likelihood Gaussian is monotonically increasing on all values that lie to the left of its mean and prior is a constant. Since the posterior PDF is monotonically increasing on its range, the MAP estimate would be the **right limit** of the uniform prior.
Similarly, if the sample mean of likelihood Gaussian PDF is lesser than the left limit of the uniform prior, then the posterior PDF is monotonically decreasing on the range of uniform prior. This is because the likelihood Gaussian is monotonically decreasing on all values that lie to the right of its mean and prior is a constant. Since the posterior PDF is monotonically decreasing on its range, the MAP estimate would be the **left limit** of the uniform prior.

2. Interpretation:

   - We see that the error is relatively higher for the ML estimate, and the two MAP estimates are close in their error values, for small sample sizes.
   - As N increases, errors for all three estimates tend to converge to zero, and the ratio of each MAP estimate with the ML estimate also tends to 1, because for large N, the data dominates the prior and the prior gets ignored, and the MAP estimate converges to the ML estimate.
   - For small N, we would prefer the **first MAP estimate**, as for small deviations from the true mean ($\sim 9.5 - 10.5$), the Gaussian prior assigns larger probabilities to them compared to uniform prior, which gives equal probability of occurrence to even large deviations from the true mean ($\sim 11 - 11.5$). Since Gaussian prior assigns smaller probabilities to values that have a larger deviation from the prior mean, even though the error is considerable, it gets weighed or scaled down due to the small probability assigned to it whereas for

a uniform prior, the large error due to values having large deviations from the true mean doesn't get scaled down, thus, giving comparatively larger errors and poorer performance than MAP estimate 1.
For large N, all three estimates give similar performance, hence to reduce computations we will prefer using the **ML estimate**.

Problem 2

1. Since data x is drawn from $U[0,1]$ (uniform distribution) and the transformed variable y is given by $y = (-1/\lambda)log(x)$ with $\lambda = 5$, and let the sample mean be $\bar{y}$. By transformation of random variables, the analytical form of distribution of y which is $P_Y(y)$ comes out as:

$$\begin{aligned} y &= g(x) = (-1/\lambda)log(x) \\ \implies x &= g^{-1}(y) = e^{-\lambda y} \\ P_Y(y) &= P_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| \\ \because P_X(x) &= \begin{Bmatrix} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{Bmatrix} \\ \left|\frac{d}{dy}g^{-1}(y)\right| &= \lambda e^{-\lambda y} \\ \therefore P_Y(y) &= \lambda e^{-\lambda y} \ , \ y > 0 \end{aligned}$$

Thus, y has exponential distribution with parameter λ.
For an exponential distribution $Y \sim Exponential(\lambda)$, the MLE for parameter λ is:

$$\text{log likelihood function} = Nlog(\lambda) - \lambda \sum_i^N y_i$$

$$\text{So, MLE } = \hat{\lambda}^{ML} = \frac{N}{\sum_i^N y_i}$$

Now we calculate the posterior PDF.
The prior can be written as: $\gamma(\lambda; \alpha, \beta) = \dfrac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$
The likelihood for a sample $\{y_i\}$ of size N with sample mean $\bar{y}$ is: $\lambda^N e^{-N\bar{y}\lambda}$
We can calculate the evidence as:

$$\begin{aligned} \int_0^\infty P_Y(y)\gamma(\lambda; \alpha, \beta)d\gamma &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda^{\alpha+N-1} e^{-(\beta+N\bar{y})\gamma} d\gamma \\ & [\text{Substitute } x = (\beta + N\bar{y})\gamma] \\ &= \frac{\beta^\alpha}{(\beta+N\bar{y})^{\alpha+N}\Gamma(\alpha)} \int_0^\infty x^{\alpha+N-1} e^{-\beta x} dx \\ &= \frac{\beta^\alpha}{(\beta+N\bar{y})^{\alpha+N}} \frac{\Gamma(\alpha+N)}{\Gamma(\alpha)} \end{aligned}$$

Thus, the posterior can be calculated as:

$$\begin{aligned}
\frac{\text{Prior}\cdot\text{Likelihood}}{\text{Evidence}} &= \frac{\dfrac{\beta^\alpha\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}\cdot\lambda^N e^{-N\bar{y}\lambda}}{\dfrac{\beta^\alpha}{(\beta+N\bar{y})^{\alpha+N}}\dfrac{\Gamma(\alpha+N)}{\Gamma(\alpha)}} \\
&= \frac{\beta^{\alpha+N\bar{y}}\lambda^{\alpha+N-1}e^{-(\beta+N\bar{y})\lambda}}{\Gamma(\alpha+N)} \\
&= \gamma(\lambda;\alpha+N,\beta+N\bar{y})
\end{aligned}$$

For a general gamma function $\gamma(\lambda;\alpha,\beta)$, the mean is:

$$\begin{aligned}
\hat{\lambda} &= \int_0^\infty \lambda\gamma(\lambda;\alpha,\beta)d\lambda \\
&= \int_0^\infty \frac{\beta^\alpha\lambda^\alpha e^{-\beta\lambda}}{\Gamma(\alpha)}d\lambda \\
&\quad [\text{Substitute } x=\beta\gamma] \\
&= \frac{1}{\beta\Gamma(\alpha)}\int_0^\infty x^\alpha e^{-x}dx \\
&= \frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha)} \\
&= \frac{\alpha}{\beta}
\end{aligned}$$

Hence for the given posterior, its mean will be:

$$\hat{\lambda}^{\text{PosteriorMean}} = \frac{\alpha+N}{\beta+N\bar{y}}$$

2. Plot of the relative errors for both estimates:

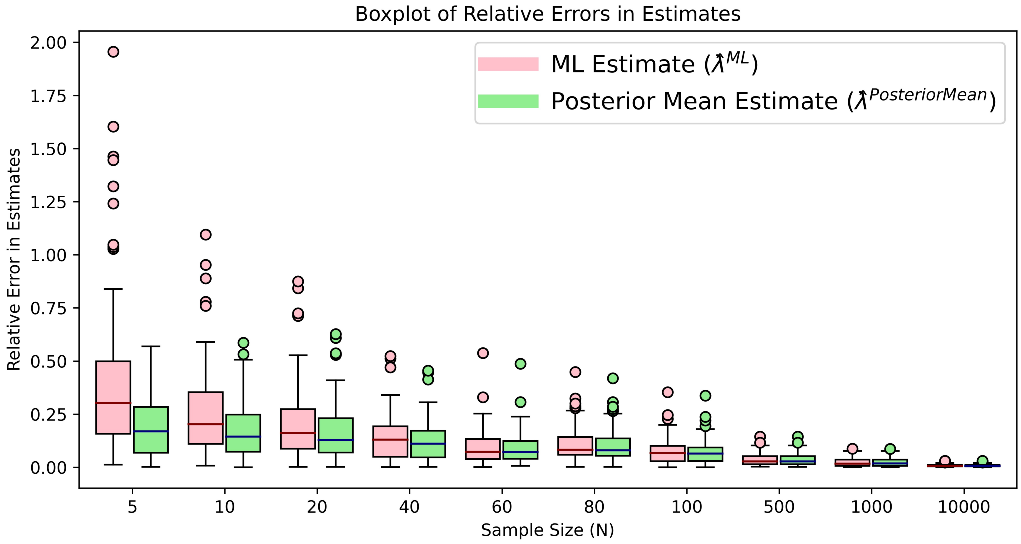

Figure 2: Relative errors for ML estimate, and Posterior Mean Estimate

3. Interpretation:
   - We see that the error is relatively higher for the ML estimate as compared to the Posterior Mean estimate, for small sample sizes.
   - As N increases, errors for both estimates converge to zero, and the ratio of the Posterior Mean estimate with the ML estimate also converges to 1, because for large N, the data dominates the prior and the prior gets ignored, and hence the Posterior Mean estimate converges to $1/\bar{y}$ which is also the ML estimate.
   - For small N, we would prefer the **Posterior Mean estimate**, as the prior assigns a large probability to the true mean value, hence even if the sample data is a little skewed and the sample mean deviates from the true mean, the prior pulls the estimated value towards the true mean, thus reducing the relative error.

     For large N, both estimates give similar performance, hence to reduce computations we will prefer using the **ML estimate**.

Problem 3

1. Maximum likelihood estimate for θ in $X \sim$ Uniform $[0,\theta]$ is $\max\{x_i\}$ where $\{x_i\}_{i=1}^N$ is the sample data.

$$\text{Likelihood function} = \left\{ \begin{array}{ll} 0 & \theta < \max\{x_i\} \\ \frac{1}{\theta^N} & \theta \geq \max\{x_i\} \end{array} \right\}$$

Derivative of likelihood function is always non-positive, so likelihood function is a non-increasing function. Therefore, maximum likelihood estimate $\hat{\theta}^{ML}$ occurs at smallest value of θ at which likelihood function is non-zero.

$$\hat{\theta}^{ML} = \max\{x_i\}$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$P(\theta|x) \propto \left\{ \begin{array}{ll} \dfrac{\theta_m^\alpha}{\theta^{\alpha+N}} & \theta_m \leq \theta \;\; \& \;\; \max\{x_i\} \leq \theta \\ 0 & \text{otherwise} \end{array} \right\}$$

$$P(\theta|x) \propto \left\{ \begin{array}{ll} \dfrac{\theta_m^\alpha}{\theta^{\alpha+N}} & \max\{\theta_m, \{x_i\}\} \leq \theta \\ 0 & \text{otherwise} \end{array} \right\}$$

$$P(\theta|x) = \text{Pareto}(\theta; \text{scale} = \max\{\theta_m, \{x_i\}\}, \text{shape} = \alpha + N)$$

For MAP estimate, we find derivative of posterior PDF (assuming constant of proportionality is positive). Let scale parameter of posterior be denoted by β.

$$\frac{d}{d\theta}P(\theta|x) \propto \frac{-(\alpha+N)\beta^\alpha}{\theta^{\alpha+N+1}}$$

$$\frac{d}{d\theta}P(\theta|x) < 0 \quad \forall\theta \geq \beta$$

For $\theta < \beta$ PDF is 0, so that can't be MAP estimate as there exists a positive value of the PDF. Since derivative of posterior PDF is always negative, it is monotonically decreasing. Therefore, the MAP estimate $\hat{\theta}^{MAP}$ is the smallest value of θ for which PDF is positive.

$$\hat{\theta}^{MAP} = \beta = \max\{\theta_m, \max\{x_i\}\}$$

2. As sample size tends to infinity, $\max\{x_i\}$ tends to θ_{true} which is the ML estimate. This is because, for any $\epsilon > 0$, we have

$$\begin{aligned}\lim_{N\to\infty} P(\theta_{true} - \max\{x_i\} > \epsilon) &= \lim_{N\to\infty} \left(\frac{\theta_{true} - \epsilon}{\theta_{true}}\right)^N \\ &= 0\end{aligned}$$

Hence $\lim_{N\to\infty} \max\{x_i\} = \theta_{true}$, and $\lim_{N\to\infty} \hat{\theta}^{ML} = \lim_{N\to\infty} \max\{x_i\} = \theta_{true}$.
Case 1: $\theta_{true} \geq \theta_m$, so (here θ_{true} is true value of parameter θ)

$$\lim_{N\to\infty} \hat{\theta}^{MAP} = \lim_{N\to\infty} \max\{\theta_m, \max\{x_i\}\} = \max\{\theta_m, \theta_{true}\} = \theta_{true}$$

Here, since both of them converge to the same value θ_{true}, we can say that $\hat{\theta}^{MAP}$ converges/tends to $\hat{\theta}^{ML}$ as the sample size tends to infinity.
Case 2: $\theta_{true} < \theta_m$, so

$$\lim_{N\to\infty} \hat{\theta}^{MAP} = \lim_{N\to\infty} \max\{\theta_m, \max\{x_i\}\} = \max\{\theta_m, \theta_{true}\} = \theta_m$$

Here, since both of them converge to different, distinct values, we can say that $\hat{\theta}^{MAP}$ does not tend to $\hat{\theta}^{ML}$ as the sample size tends to infinity.

Case 1 is **desirable** as whenever the MAP estimate converges to the ML estimate $\hat{\theta}^{ML}$, it is an indication that we have chosen a good prior, and that both of them will converge towards the true value, as **MLE always converges to the true value** θ_{true} in the limit of large sample size, and hence the MAP estimator is **consistent**.
Case 2 is **not desirable** as the MAP estimate doesn't converge to the MLE, so the MAP estimate doesn't give us the true value of the parameter in the limit of large sample size leading to an unsatisfactory estimate which is an indication that we haven't chosen a good prior.

3. For a Pareto distribution $P(\theta) = c(\theta_m/\theta)^\alpha$ when $\theta \geq \theta_m$, and $P(\theta) = 0$ otherwise, where c is the constant of proportionality. We calculate this c first:

$$\begin{aligned}1 &= \int_{-\infty}^{\infty} P(\theta) \\ &= \int_{\theta_m}^{\infty} c(\theta_m/\theta)^\alpha \\ &= c\ \theta_m^\alpha \left(\frac{\theta^{-\alpha+1}}{-\alpha+1}\Bigg|_{\theta_m}^{\infty}\right)\end{aligned}$$

Since $\alpha > 1$, we conclude that $c = \dfrac{\alpha - 1}{\theta_m}$.
For mean, E[θ] of the Pareto distribution $P(\theta)$:

$$\begin{aligned}
\mathbf{E}[\theta] &= \int_{-\infty}^{\infty} \theta P(\theta) \\
&= \int_{\theta_m}^{\infty} \theta P(\theta) \\
&= \int_{\theta_m}^{\infty} c \frac{\theta_m^{\alpha}}{\theta^{\alpha-1}} \\
&= c\theta_m^{\alpha} \frac{\theta^{-\alpha+2}}{-\alpha+2}\bigg|_{\theta_m}^{\infty} \\
\mathrm{E}[\theta] &= \left\{ \begin{array}{cc} \dfrac{(\alpha-1)\theta_m}{\alpha-2} & \alpha > 2 \\ +\infty & \alpha \leq 2 \end{array} \right\}
\end{aligned}$$

So, for our Pareto posterior distribution:

$$P(\theta|x) = \text{Pareto}(\text{scale}(\beta) = \max\{\theta_m, \{x_i\}\}, \text{shape} = \alpha + N)$$

Therefore the Posterior mean estimate $\hat{\theta}^{\text{PosteriorMean}} = \dfrac{(\alpha + N - 1)\beta}{\alpha + N - 2}$, since $\alpha > 1$ and N can be assumed to be ≥ 1 as we will have at least 1 sample data point.

4. $\hat{\theta}^{\text{PosteriorMean}}$ is similar to $\hat{\theta}^{\text{MAP}}$ in the limit of large sample size, because

$$\begin{aligned}
\lim_{N\to\infty} \hat{\theta}^{\text{PosteriorMean}} &= \lim_{N\to\infty} \frac{\alpha + N - 1}{\alpha + N - 2} \max\{\theta_m, \max\{x_i\}\} \\
&= \lim_{N\to\infty} \frac{1 + (\alpha - 1)/N}{1 + (\alpha - 2)/N} \max\{\theta_m, \max\{x_i\}\} \\
&= \lim_{N\to\infty} \max\{\theta_m, \max\{x_i\}\} \\
&= \lim_{N\to\infty} \hat{\theta}^{\text{MAP}}
\end{aligned}$$

So, $\hat{\theta}^{\text{PosteriorMean}}$ tends to $\hat{\theta}^{\text{MAP}}$ as the sample size tends to infinity. Again, it implies that **Case 1** ($\theta_{true} \geq \theta_m$, and correspondingly $\hat{\theta}^{\text{PosteriorMean}}$ converges to $\hat{\theta}^{\text{ML}}$) is desirable as it indicates we have chosen a suitable prior, and the estimate will infact converge to the true value of θ (i.e. the Posterior Mean estimator is **consistent**) because ML estimate $\theta^{\hat{ML}}$ converges to the true value θ_{true}.
For **Case 2** ($\theta_{true} < \theta_m$) the MAP estimate doesn't converge to the MLE, so the posterior mean estimate doesn't converge to the MLE and thereby does not give us the true value of the parameter which is an undesirable result.