CS 215 : Data Analysis and Interpretation
Assignment : Bayesian Estimation

Instructor : Suyash P. Awate

Submission Instructions:

• IITB and CSE have zero tolerance to plagiarism.

• For the sake of effective learning, if you submit the solution to the assignment as a group, then each member of the group agrees to have participated fully (100%) in performing every part of every question in the assignment.

• If you submit the solution to the assignment (by yourself or as a group), then you agree that every line of code and every line in the report is your (or your group's) own, and isn't a copied/modified version of any other source online (on the internet) or offline (in electronic form or paper form or any other form).

• If you submit the solution to the assignment as a group and any member of the group is determined to have committed any (non-zero amount of) plagiarism, then the full penalty (a reduction of at least 1 letter grade, e.g., from AB to BB or even lower) will be applicable to every member of the group. The penalty will be applicable to givers and takers both.

• Submit your solution to each problem, i.e., (i) the code, (ii) the results, e.g., graphs or other data, and (iii) the report (in Adobe PDF format), for each question, through moodle. Put the code within the folder "code", the results within the folder "results", and the report within the folder "report".

• Submit all code that allows the TAs to regenerate your results, exactly as they appear in the report.

• Submit a single zip file that contains the solutions to all problems in the assignment.

• To get any possible partial credit for the code, ensure that the code is very well documented. To get partial credit for the derivations, include all derivation steps in their full details.

• To avoid non-deterministic results in each program run, and to make the results reproducible during test time, use rng(seed) where seed is a fixed hard-coded integer in your code.

• If the question suggests the use of some function in Matlab, then you can use a corresponding function in other coding frameworks/languages.

• Delayed submissions will be penalized 25% of the total points on each day after the deadline, i.e., submitting anytime within the first 24 hours after the deadline will incur a penalty of 25% of the total

points.

• If you feel there is a typo in the question, please make suitable assumptions, consistent with those in the question, and proceed to solve the problem. Also, in that case, please let the TAs or the instructor know.

• **5 points** are reserved for submission in the proper format.

1. (10 points) Use the Matlab function randn() to generate a data sample of N points drawn from a Gaussian distribution with mean $\mu_{\text{true}} = 10$ and standard deviation $\sigma_{\text{true}} = 4$. Consider the problem of using the data to get an estimate $\widehat{\mu}$ of this Gaussian mean, assuming it is unknown, when the standard deviation σ_{true} is known.

   Consider using one of the two prior prior distributions on the mean: (i) a Gaussian prior with mean $\mu_{\text{prior}} = 10.5$ and standard deviation $\sigma_{\text{prior}} = 1$ and (ii) a uniform prior over $[9.5, 11.5]$.

   Consider various sample sizes $N =$ 5, 10, 20, 40, 60, 80, 100, 500, $10^3, 10^4$. For each sample size N, repeat the following experiment $M \geq 100$ times: generate the data, get the maximum likelihood estimate $\widehat{\mu}^{\text{ML}}$, get the maximum-a-posteriori estimates $\widehat{\mu}^{\text{MAP1}}$ and $\widehat{\mu}^{\text{MAP2}}$, and measure the relative errors $|\widehat{\mu} - \mu_{\text{true}}|/\mu_{\text{true}}$ for all three estimates.

   • (8 points) Plot a single graph that shows the relative errors for each value of N as a box plot (use the Matlab boxplot() function), for each of the three estimates.

   • (2 points) Interpret what you see in the graph. (i) What happens to the error as N increases ? (ii) Which of the three estimates will you prefer and why ?

2. (15 points) Use the Matlab function rand() to generate a data sample of N points from the uniform distribution on $[0, 1]$. Transform the resulting data x to generate a transformed data sample where each datum $y := (-1/\lambda)\log(x)$ with $\lambda = 5$. The transformed data y will have some distribution with parameter λ; what is its analytical form ? Use a Gamma prior on the parameter λ, where the Gamma distribution has parameters $\alpha = 5.5$ and $\beta = 1$.

   Consider various sample sizes $N =$ 5, 10, 20, 40, 60, 80, 100, 500, $10^3, 10^4$. For each sample size N, repeat the following experiment $M \geq 100$ times: generate the data, get the maximum likelihood estimate $\widehat{\lambda}^{\text{ML}}$, get the Bayesian estimate as the posterior mean $\widehat{\lambda}^{\text{PosteriorMean}}$, and measure the relative errors $|\widehat{\lambda} - \lambda_{\text{true}}|/\lambda_{\text{true}}$ for both the estimates.

   • (5 points) Derive a formula for the posterior mean.

   • (8 points) Plot a single graph that shows the relative errors for each value of N as a box plot (use the Matlab boxplot() function), for both the estimates.

   • (2 points) Interpret what you see in the graph. (i) What happens to the error as N increases ? (ii) Which of the two estimates will you prefer and why ?

3. (20 points) Suppose random variable X has a uniform distribution over $[0, \theta]$, where the parameter θ is unknown. Consider a Pareto distribution prior on θ, with a scale parameter $\theta_m > 0$ and a shape parameter $\alpha > 1$, as $P(\theta) \propto (\theta_m/\theta)^\alpha$ for $\theta \geq \theta_m$ and $P(\theta) = 0$ otherwise.

   • (5 points) Find the maximum-likelihood estimate $\widehat{\theta}^{\text{ML}}$ and the maximum-a-posteriori estimate $\widehat{\theta}^{\text{MAP}}$.

- (8 points) Does $\widehat{\theta}^{\text{MAP}}$ tend to $\widehat{\theta}^{\text{ML}}$ as the sample size tends to infinity ? Is this desirable or not ?
- (5 points) Find an estimator of the mean of the posterior distribution $\widehat{\theta}^{\text{PosteriorMean}}$.
- (2 points) Does $\widehat{\theta}^{\text{PosteriorMean}}$ tend to $\widehat{\theta}^{\text{ML}}$ as the sample size tends to infinity ? Is this desirable or not ?