STAT 565, Assignment 4

assignment due 12/12

Name 1 (First, Last) ______________________

Name 2 (First, Last) ______________________

Name 2 (First, Last) ______________________

Instructions:

- This is an open book, notes, computer, R, and internet assignment. This assignment is team work, teams as announced by email or in class. No communication with other persons –except me– about the assignment is allowed, including communications with real persons over internet. Submit only your team's work. If you need clarification about questions please let me know. It is okay to ask questions.
- Please see the syllabus for submission details on take-home assignments. Only one copy per team is needed. All team members should make reasonable contribution to the work. Solutions to problems in Theory section are best submitted as pdf. Solutions to problems in Practice section are to be submitted only as R programs. Please write team members names at the top of your scripts and comment your code appropriately so that it is easy to follow.
- Solutions to problems in Theory part need solid algorithmic or mathematical justifications to get full credit. No simulation solutions will be accepted unless the problem specifies otherwise. However, you can use simulations to gain insight to the problem. Use precise mathematical notation, paying particular attention to clearly denote random variables and fixed values. Define any quantity used in the solution but not given in the problem.
- Solutions to problems in Practice part need a working R program that produces a clear answer for each part of the problem. For long problems, writing

multiple R files in the form of R functions and bundling them in a main.R might be a good idea. Each problem needs at least one separate R file. When submitting by email, please send compressed folder (.zip) for programs, but do not include any output. Programs should run to produce the output.

- All questions within the Theory part have equal value. All questions within the Pratice part have equal value. Theory and Practice parts also have equal value (50% each in this assignment).
- Feel free to use built-in R functions for simulating random varaibles from non-standard distributions. Feel also free to use any mathematical function in R including functions for calculating combinations, permutations, Gamma function, Beta function, logarithms, and trigonometric functions.

Part I: Theory

1. We let $(X_1, X_2, \cdots, X_n)$ be an iid sample from an exponential distribution with pdf $f(x|\lambda) = \lambda e^{-\lambda x},\ x \geq 0,\ \lambda > 0$. The statistic $1/\bar{X}$ is sufficient for λ. For purposes of explaining how ABC works to your colleagues, use $1/\bar{X}$ as a statistic and write an ABC algorithm to sample the posterior distribution of parameter λ using an independent Metropolis sampler and the idea of MCMC without likelihoods. (Hint: You can probably figure this out yourself, essentially modifying the ABC by rejection that we have covered in class by replacing the rejection with independent Metropolis sampler accordingly. However, I recommend searching: “Markov chain Monte Carlo without likelihoods” by Marjoram et al 2003 PNAS article (this is a very accesible paper if you are interested in likelihood-free methods) and getting some help from the algorithms described there.)

Part II: Practice

1. For this problem please refer to the background given in *Inference with Approximate Bayesian Computation (ABC) by Rejection, (Research Level Example)* that we have covered in class. We are given the sequence:

```
OPIEHAKLESSXFYHBWWEEFLKMRFJAIICNSAIFNDFAJNJVSEUFBEHHRFEQUAQ
```

in which our ciphered message (single word) sits. The observed key, after five interference events acted on the true key is given in `ABCPractice.RData` file which contains a 26×64 matrix of 1s and 0s where the sum of each row is the score n_ℓ for each letter of the English alphabet, starting with `A` and ending with `Z`.

This problem is not only about testing a computational statistical method but also about model building.

a. First, we will consider an early stage of model building process and will test a very straightforward case with ABC. Assume that for each interference, the probabilities are known and quite extreme:

$$\theta^{(1)} = 0.01,\ \theta^{(2)} = 0.99,\ \theta^{(3)} = 0.99, \theta^{(4)} = 0.97,\ \theta^{(5)} = 0.02.$$

Implement the ABC by rejection to sample the posterior distribution of n_ℓ for each letter. You can run the program independently for all letters. Note that the observed score is a discrete statistic since it takes values on $\{0, 1, \cdots, 64\}$. So we can set the (tuning) tolerance parameter ϵ which has to be positive in ABC when the statistic is continuous to zero. This means that one error source in ABC can be eliminated in this case. However, if the acceptance rate is very low because you try to match the observed and simulated values of the statistic exactly, then you can increase ϵ, although this will make harder to perform the correct inference. Take the posterior mode as the estimate of n_ℓ. If it is greater than 32 consider the letter as part of the message, else discard the letter in the sequence. Report the deciphered message as your answer.

b. As the second stage in model building we will assume $\theta^{(t)}$ are unknown and try to sample the joint posterior distribution of n_ℓ and $\theta^{(t)}$. First, we want to add some but not too much variability in $\theta^{(t)}$ and see if we still can recover the message (probably not without any additional methods employed). Assume $\theta^{(t)}$ are unknown and assign them a beta prior (natural choice since $\theta \in (0, 1)$) with small variance (you can assume prior independence). Make the variance *very small* at first with mode of the beta distribution situated at the given values of θ as above, obtain the sample from the joint posterior and see if you can still recover the message using estimated $n\ell$. If the answer is yes, then increase the variance a bit and try again. After some trials report a ballpark value of the prior variance when we cannot recover the message reasonably (we can easily formalize this by building a systematic setup in which variance is incrementally increased, but no need for our purposes).

The mode of the beta distribution is given by

$$\frac{\alpha + \beta - 1}{\alpha + \beta - 2}$$

and the variance is given by

$$\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

so that you can choose $\alpha, \ \beta$ accordingly.