MATH70093 – Computational Statistics; Assessed coursework 2—Autumn 2021
Due Monday 22nd of November 2021– deadline specific to your time zone

Upload your final version only - once the report is uploaded there is no option for re-uploading. Avoid last minute uploads. **Hand-in no more than <u>10</u> pages.** Provide your `R` code in the appendix; the appendix does not count towards the page limit. Considerable emphasis will be put on clarity of expression and a clean presentation *(5 marks allocated to presentation)*. Only detailed, well-written answers that clearly explain your reasoning will score highly.

1. *(25 marks)* Consider the density f given by $f(x) = k/(1 + |x - 2|^3)$ for $x \in [0, 5]$ and 0 otherwise; where $k > 0$ is some normalising constant. Construct a Metropolis-Hastings sampler to sample from this density f using a random walk proposal with additive noise following a normal distribution with standard deviation σ.

   (a) For $\sigma = 1$, run the sampler for 5000 iterations and estimate $E(X^3)$ and $P(X < 1)$, assuming X is a random variable with density f. Check your estimate by comparing it to the result that you obtain using a numerical integration method of your choice.

   (b) For $\sigma = 0.05$, create four independent Markov chains by running the Metropolis-Hastings sampler for 5000 iterations for each run starting from different values, e.g. 1, 2, 3 and 4. Compare the empirical cdf of the four resulting samples to a plot of the true cdf (which you can get using numerical integration). Discuss.

   (c) Compare the efficiency of the Metropolis-Hasting sampler for $\sigma \in \{0.05, 1, 5\}$ in terms of the acceptance rate and of the resulting stationary distributions.

2. *(25 marks)* Suppose we observe the following iid sample from a random variable X

   -0.71, -1.30, -0.13, -2.03, 1.62, 2.38, 0.48, 0.51, -0.69, -2.32, -2.02, 1.23, -0.25, 0.76, 0.65
   [check: 15 numbers with mean approximately equal to -0.1213]

   (a) Using a non-parametric bootstrap with 1000 bootstrap samples, compute standard and studentized 95% two-sided confidence intervals for $E[\cos(X)]$.

   (b) We want to examine the coverage probability of these confidence intervals. For this, assume a true data generating mechanism that $X_1, \ldots X_n$ follows independently a Student's t distribution with 4 degrees of freedom. Generate 500 new datasets of length $n = 15$ and construct the corresponding confidence intervals using 1000 bootstrap samples each. Compare the coverage probability for the two types of intervals.

   (c) Suppose we are now interested in $T(X_1 \ldots X_n) = \text{median}(\cos(X_1) \ldots \cos(X_n))$. Using a non-parametric bootstrap with 1000 bootstrap samples, compute a standard 95% two-sided confidence intervals for $T(X_1 \ldots X_n)$. Describe with mathematical equations an approach to compute a studentised 95% confidence interval or $T(X_1 \ldots X_n)$. Implement the approach using 500 bootstrap samples.

3. *(20 marks)* Denote the last 2 digits of your CID as a_1 and a_2, e.g, if your CID is 00268914, then take $a_1 = 1$ and $a_2 = 4$. Construct a Gibbs sampler to sample from the distribution with density f given by

   $$f(x, y) = k\, y^{1/2} \exp\left\{-\frac{x^2}{2} - \frac{y}{2}\left(1 + (a_1 - x)^2 + (a_2 - x)^2\right)\right\} \text{ for } y \geq 0$$

   where $k > 0$ is a normalising constant. Describe the sampler with mathematical equations. Run the sampler for 2000 iterations and produce a scatter plot of the approximate sample from the joint distribution. Estimate $P((X, Y) \in [-1, 1] \times [0, 0.5])$ and $E(X^2)$.

4. *(25 marks)* Consider the dataset in the file `qu4.csv` on Blackboard containing observations from two different groups (50 observations in group 1 and 75 in group 2). We are interested in testing the null hypothesis that the two groups have the same variance.

   (a) Devise a permutation test using the absolute difference of the sample variances as a test statistic. Report the p-value and the test decision at a 5% level.

   (b) Assume now that the data have been sampled independently from Gamma distributions with shape parameter $k = 5$ and unknown scale parameters θ_1 and θ_2 respectively for group 1 and 2.

      i. Use parametric bootstrap to estimate the distributions of the sample variance of each group. Precisely describe the approach. Plot the empirical cdf of the two resulting distributions. *[Hint: the scale parameter θ_1 can be estimated by the empirical mean of the data in group 1 divided by k.]*

      ii. Devise a parametric bootstrap test by using parametric bootstrap to estimate the distribution of the absolute difference of the sample variances under the null hypothesis. Precisely describe the approach Report the p-value and the decision at a 5% level. Discuss the result in comparison to question a).