COMPUTER SCIENCE TRIPOS Part IB – mock – Paper 6

2 Foundations of Data Science (DJW)

Let $x_1, \ldots, x_n$ be observed values, which we believe are sampled independently from the distribution Uniform$[\mu - \theta, \mu + \theta]$, for some parameters $\mu \in \mathbb{R}$ and $\theta > 0$.

(*a*) Suppose μ is known and θ is unknown. Use $\Theta \sim \text{Pareto}(b_0, \alpha_0)$ as the prior for θ, where b_0 and α_0 are constants. (The Pareto distribution is described below.)

(*i*) What is the prior density of Θ? [1 mark]

(*ii*) Show that the posterior distribution of Θ is Pareto, and give its parameters. [5 marks]

(*iii*) Calculate a 95% posterior confidence interval for Θ. [4 marks]

(*b*) Suppose μ and θ are both unknown. Use Normal(c_0, σ_0^2) as the prior for μ, and Pareto(b_0, α_0) as the prior for θ. Here c_0, σ_0, b_0, and α_0 are all constants.

(*i*) Find the joint posterior density of the two parameters. [*Note:* Leave your answer as an unnormalized density function.] [3 marks]

(*ii*) Give pseudocode to generate a weighted sample from this density. Your code should produce a list of m sampled pairs $\big[(\mu_1, \theta_1), \ldots, (\mu_m, \theta_m)\big]$ together with weights $[w_1, \ldots, w_m]$. [3 marks]

(*iii*) Give pseudocode to find a 95% posterior confidence interval for Θ. [4 marks]

Note: If $X \sim \text{Pareto}(b, \alpha)$ then it has cumulative distribution function

$$\mathbb{P}(X \leq x) = \left\{1 - \left(\frac{b}{x}\right)^{\alpha}\right\} 1_{x \geq b}$$

and it may be sampled using

$$\mathsf{b * (1 + numpy.random.pareto(a=}\alpha\mathsf{))}$$

Answer: Part $(a)(i)$. Simply differentiate the cdf to get

$$\Pr_\Theta(\theta) = \alpha_0 \left(\frac{b_0^{\alpha_0}}{\theta^{\alpha_0+1}} \right) 1_{\theta \geq b_0}.$$

Or, if the indicator function seems a bit mysterious, write out the cases separately:

$$\mathbb{P}(\Theta \leq \theta) = \begin{cases} 1-(b_0/\theta)^{\alpha_0} & \text{if } \theta \geq b_0 \\ 0 & \text{if } \theta < b_0 \end{cases}$$

$$\Pr_\Theta(\theta) = \frac{d}{d\theta}\mathbb{P}(\Theta \leq \theta) = \begin{cases} \alpha_0 b_0^{\alpha_0}/\theta^{\alpha_0+1} & \text{if } \theta \geq b_0 \\ 0 & \text{if } \theta < b_0. \end{cases}$$

I don't want to see any x left in the answer! When you transpose equations from one setting to another, make sure you transpose the symbols. Here we're transposing from X to Θ, from x to θ, from b to b_0, from α to α_0.

Part $(a)(ii)$. The density for the entire dataset is

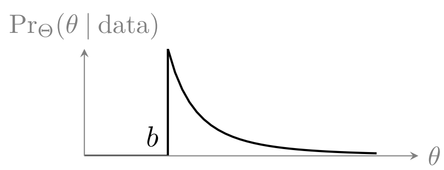
$$\begin{aligned} \Pr\big(x_1,\dots,x_n \mid \theta\big) &= \prod_{i=1}^n \Pr_X(x_i \mid \theta) \quad \text{where } X \sim \text{Uniform}[\mu-\theta,\mu+\theta] \\ &= \prod_{i=1}^n \left\{ \frac{1}{2\theta} 1_{x_i \in [\mu-\theta,\mu+\theta]} \right\} \\ &= \frac{1}{(2\theta)^n} 1_{\text{all } x_i \in [\mu-\theta,\mu+\theta]} \\ &= \frac{1}{(2\theta)^n} 1_{\min_i x_i \geq \mu-\theta} 1_{\max_i x_i \leq \mu+\theta}. \end{aligned}$$

The density depends on μ too, of course, but because the question tells us that μ is known I'm not bothering to write it out. See also example sheet 1 question 5 for practice at working with indicator functions in this way. So the posterior density is

$$\begin{aligned} \Pr_\Theta(\theta \mid x_1,\dots,x_n) &= \kappa \Pr_\Theta(\theta) \Pr(x_1,\dots,x_n \mid \mu,\theta) \\ &= \kappa \alpha_0 b_0^{\alpha_0} \frac{1}{\theta^{\alpha_0+1}} \frac{1}{(2\theta)^n} 1_{\theta \geq \mu - \min_i x_i} 1_{\theta \geq \max_i x_i - \mu} \\ &= \kappa' \frac{1}{\theta^{\alpha_0+n+1}} 1_{\theta \geq \max\{b_0, \mu-\min_i x_i, \max_i x_i - \mu\}}. \end{aligned}$$

We only care about how this depends on θ, not about the constants, so we can amalgamate all non-θ terms into κ'. (To be precise, $\kappa' = \kappa\alpha_0 b_0^{\alpha_0}/2^n$.) The constants κ and κ' are whatever they need to be in order for this posterior density to integrate to 1. In fact, we don't need to even work out κ'—we see that the θ terms are exactly that of a Pareto distribution with parameters $b = \max\{b_0, \mu - \min_i x_i, \max_i x_i - \mu\}$ and $\alpha = \alpha_0 + n$, and so κ' must necessarily be precisely the constant term at the front of the Pareto density. In conclusion, the posterior distribution is $(\Theta \mid x_1,\dots,x_n) \sim \text{Pareto}(b,\alpha)$ where b and α are as stated.

Part $(a)(iii)$. In many situations it's reasonable to report a two-sided confidence interval, but in this case, after sketching the posterior density, a one-sided confidence interval seems more appealling. The posterior density looks like this:



So let's report a one-sided confidence interval $[b, \mathsf{hi}]$. We need to choose hi so that $\mathbb{P}(\Theta > \mathsf{hi} \,|\, \text{data}) = 0.05$.

We know that $(\Theta|\text{data})$ is Pareto(b, α) with parameters b and α as given in the answer to part $(a)(ii)$. Using the cdf as stated in the question,

$$\mathbb{P}(\Theta > \mathsf{hi} \mid \text{data}) = \begin{cases} (b/\mathsf{hi})^\alpha & \text{if } \mathsf{hi} \geq b \\ 1 & \text{if } \mathsf{hi} < b \end{cases}$$

and so we pick hi to solve $(b/\mathsf{hi})^\alpha = 0.05$ hence $\mathsf{hi} = b(0.05)^{-1/\alpha}$.

Part $(b)(i)$. Bayes's rule is always the same: posterior density = constant, times prior density, times data density. When we apply Bayes's rule we MUST include all unknown parameters in the prior and posterior. In this case there are two unknowns, μ and θ, so we'll write down a joint prior.

Assuming the prior parameters M and Θ are independent and noting that M is the upper case version of μ!, the prior density is

$$\Pr_{(M,\Theta)}(\mu, \theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(\mu - c_0)^2/2\sigma_0^2} \left(\frac{b_0^{\alpha_0}}{\theta^{\alpha_0+1}} \right) 1_{\theta \geq b_0}.$$

The data density was calculated in part $(a)(ii)$: it is

$$\Pr(x_1, \ldots, x_n \mid \theta, \mu) = \text{const} \times \frac{1}{\theta^n} 1_{\min_i x_i \geq \mu - \theta} 1_{\max_i x_i \leq \mu + \theta}.$$

Putting them together, the joint posterior density is

$$\Pr_{(M,\Theta)}(\mu, \theta \mid x_1, \ldots, x_n) = \kappa e^{-(\mu - c_0)^2/2\sigma_0^2} \frac{1_{\theta \geq b_0}}{\theta^{\alpha_0+1}} \frac{1}{\theta^n} 1_{\mu - \theta \leq m} 1_{M \leq \mu + \theta}.$$

Here I've gathered non-θ terms into the constant κ. The question doesn't say "simplify your formula as much as possible", so I'll leave it like this. If I thought I'd need to use this formula to answer a later part of the question, I'd simplify it.

Part $(b)(ii)$

```
# Sample m pairs from the joint prior distribution
# For clarity, write random.pareto2(b,a) for b * (1+random.pareto(a))
prior_samp = [(random.normal(loc=c0, scale=sigma0, size=m), random.pareto2(b0,alpha0))
              for _ in range(m)]

# Compute the data density, and normalize to get weights w
n,m,M = len(x), min(x), max(x)
prx = [1/(theta ** n) * indicator(mu-theta <= m and M <= mu+theta)
       for (mu,theta) in prior_samp]
w = prx / sum(prx)
```

The question says "produce a list of sampled pairs". It says this to emphasize that we're sampling from the joint distribution. The code above does this. But in practical numpy code, it's not good style to use pairs — we could if we wanted column-stack the `theta` and `mu` vectors, or we could just leave them unstacked, i.e. just generate two separate vectors for `theta` and `mu`. That will still get you the marks.

Part $(b)(iii)$ We're asked for a 95% posterior confidence interval for Θ. Our sample actually consists of pairs (M, Θ), so we might start by just concentrating on the Θ samples:

```
theta_samp = [theta for (mu,theta) in prior_samp]
```

Recall the weighted Monte Carlo approximation:

$$\mathbb{P}(\Theta \in A \mid \text{data}) \approx \sum_{i:\theta_i \in A} w_i.$$

We want to pick an interval $[\mathsf{lo}, \mathsf{hi}]$ that has probability 95%, i.e. so that the total weight of samples inside this interval is 95%. We can choose whatever interval we like, as long as it has the correct weight.

```
# Pick lo such that sum(w[theta<lo])=0.025
# and hi such that sum(w[theta>hi])=0.025:

(theta_samp, w) = theta_samp and w, but sorted in order of increasing theta
F = cumsum(w)
lo = the last theta for which F <= 0.025
hi = the last theta for which F <= 0.975
```

Don't worry about rounding, and which exact side of the interval we should be on. The code in lecture notes didn't bother about it, so you don't have to either. The prior distribution for θ is continuous, and if the sample is large enough, there should be no problem. It's a bit more of a nuisance to deal with discrete distributions—but again, this isn't something worth dealing with in the exam.