IMPERIAL

MATH70090

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2024

This paper is also taken for the relevant examination for the
Associateship of the Royal College of Science

Bayesian Methods

Date: Tuesday, April 30, 2024

Time: 10:00 – 11:30 (BST)

Time Allowed: 1.5 hours

This paper has 2 Questions.

Please Answer All Questions in 1 Answer Booklet

Candidates should start their solutions to each question on a new sheet of paper.

Supplementary books may only be used after the relevant main book(s) are full.

Any required additional material(s) will be provided.

Allow margins for marking.

Credit will be given for all questions attempted.

Each question carries equal weight.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

The open-book material allowed during the examinations consists of any material provided by the lecturers and annotated by the students, i.e. annotated lecture notes, annotated slides, and annotated problem class sheets. Books and electronic devices are not allowed.

1. A manufacturer of high-precision digital audio monitoring equipment has built its reputation on being able to produce very precisely-calibrated detectors. When exposed to an audio signal of (true) amplitude A the result is a measured amplitude $\hat{A}$, which is a draw from some distribution $P(\hat{A}|A, D)$, the form of which is determined by the physical properties of the detector, D.

One of the company's engineers is given the task of characterising the behaviour of the detector, which they do by making $N \gg 1$ measurements of a reference source of known amplitude, A_0, which yields measurements $\hat{A}_{1:N} = (\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_N)$. The engineer's task is then to use these data to infer the form of $P(\hat{A}|A_0, D)$.

(i) The engineer starts by making an assumption of normally distributed noise, N, such that $P(\hat{A}|A_0, D) = \mathrm{N}(\hat{A}; A_0, \sigma^2)$, where $\mathrm{N}(x; \mu, \sigma^2)$ is a normal density of mean μ and variance σ^2.

(a) Adopting an improper uniform prior for $\sigma > 0$ based on the minimal background knowledge, K, find the posterior distribution $P(\sigma|\hat{A}_{1:N}, A_0, \mathrm{N}, K)$, simplifying your answer as much as possible. Be sure to note any assumptions you make and to assess if the posterior is normalizable; but there is no requirement to calculate the normalization constant.

(b) What is the implied (predictive) probability distribution for some future measurement of this reference source, $P(\hat{A}|\hat{A}_{1:N}, A_0, \mathrm{N}, K)$? [This again can be calculated up to a normalization constant.]

(c) Is the assumption that the noise is normally distributed, N, justified on the basis of the available information? Explain your answer.

(d) This assumed model for $P(\hat{A}|A_0, D)$ implies that the device's outputs are unbiased. Is there evidence for this?

(ii) The engineer notices that there are a few outliers in the data and so considers using a scaled Student-t distribution, with an additional shape parameter which controls how heavy the tails are.

(a) Why would a heavy-tailed distribution be appropriate in the presence of outliers?

(b) Would a heavy-tailed distribution yield valid results if there were no outliers?

(c) Outline in point form how the engineer could use the test data to assess which of these two noise models is more appropriate, in particular accounting for the fact that both models have unspecified internal parameters.

(iii) The engineer subsequently decides that both the above approaches are too restrictive, and instead decides to explore a "non-parametric" model in which $P(\hat{A}|A_0, D)$ is represented as a (potentially-infinite) mixture model of normal densities with a Dirichlet Process prior on their weights.

(a) Describe the mathematical structure of this model, highlighting the full set of parameters which should be included in the inference calculation.

(b) Explain how this model can be set up to avoid over-fitting of the data.

[Total 30 marks]

2. A high-school physics class is given an assignment to measure the speed of sound, s, and the speed of light, c. They do this by timing echoes (sound) and reflections (light) from the wall of a nearby building a distance of $D = 490$ metres from their classroom. The round trip travel time for a signal travelling at speed v (here equal to either s or c) would be $T = 2\,D/v$, but it is of course impossible to measure this time perfectly. The dominant source of uncertainty – the only one which needs to be considered here – is that the students are making their timings using an old-style digital watch which only displays seconds. A measured time of, say, $\hat{T} = 14$ would only provide the constraint that $13.5\,\text{seconds} \leq T < 14.5\,\text{seconds}$, implying that $P(\hat{T}|T)$ is a uniform distribution.

(i) The students decide their only secure prior knowledge, K, is that light travels faster than sound (as, e.g., thunder is heard after lightning). They hence adopt an improper prior distribution of the form

$$P(s, c|K) \propto \Theta(s)\,\Theta(c)\,\Theta(c - s),$$

where $\Theta(x)$ is the Heaviside step function, equal to 0 if $x < 0$ and 1 if $x \geq 0$.

(a) Assess if this distribution effectively represents the stated prior knowledge and/or if it encodes additional information beyond that explicitly listed.

(b) Give two reasons why it can be useful to use an improper prior.

(c) State two potential difficulties of using an improper prior.

(ii) The students start by measuring the speed of sound, obtaining $\hat{T} = 3$.

(a) What is the implied posterior distribution, $P(s|\hat{T}, D, K)$?

(b) Does the prior distribution cause any problems? Explain your answer.

(iii) The students now turn to measuring the speed of light but the reflection appears so rapidly that they measure $\hat{T} = 0$.

(a) What is the implied posterior distribution, $P(c|\hat{T}, D, K)$?

(b) Does the prior distribution cause any problems? Explain your answer.

(c) If the students made repeat measurements and they all yielded $\hat{T} = 0$, would the implied constraints on c be improved? Explain your answer.

(iv) One of the students re-measures the distance to the building and gets a different result from that quoted above, implying some potentially significant uncertainty about the value of D. Outline in point form how this additional source of uncertainty could be properly and robustly incorporated into the data analysis.

[Do *not* attempt a full calculation.]

[Total 30 marks]

End of examination paper

Imperial College London

Module:	MATH70090
Setter:	Daniel Mortlock
Checker:	Alastair Young
Editor:	Zak Varty
External:	Dave Woods
Date:	March 11, 2024

MSc EXAMINATIONS (STATISTICS)

MATH70090 Bayesian Methods

Time: 1 hour 30 minutes

Setter's signature	Checker's signature	Editor's signature
.....	.....	.....

Questions on next page

1. A manufacturer of high-precision digital audio monitoring equipment has built its reputation on being able to produce very precisely-calibrated detectors. When exposed to an audio signal of (true) amplitude A the result is a measured amplitude $\hat{A}$, which is a draw from some distribution $P(\hat{A}|A, D)$, the form of which is determined by the physical properties of the detector, D.

One of the company's engineers is given the task of characterising the behaviour of the detector, which they do by making $N \gg 1$ measurements of a reference source of known amplitude, A_0, which yields measurements $\hat{A}_{1:N} = (\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_N)$. The engineer's task is then to use these data to infer the form of $P(\hat{A}|A_0, D)$.

(i) The engineer starts by making an assumption of normally distributed noise, N, such that $P(\hat{A}|A_0, D) = \mathrm{N}(\hat{A}; A_0, \sigma^2)$, where $\mathrm{N}(x; \mu, \sigma^2)$ is a normal density of mean μ and variance σ^2.

(a) Adopting an improper uniform prior for $\sigma > 0$ based on the minimal background knowledge, K, find the posterior distribution $P(\sigma|\hat{A}_{1:N}, A_0, \mathrm{N}, K)$, simplifying your answer as much as possible. Be sure to note any assumptions you make and to assess if the posterior is normalizable; but is no requirement to calculate the normalization constant.

ANSWER: (SIMILAR TO SEEN)
The unnormalized posterior is proportional to

$$P(\sigma|\hat{A}_{1:N}, A_0, \mathrm{N}, K) \propto P(\sigma|A_0, \mathrm{N}, K)\, P(\hat{A}_{1:N}|A_0, \sigma, \mathrm{N}, K).$$

The (improper) prior under the normal model is $P(\sigma|A_0, \mathrm{N}, K) \propto \Theta(\sigma)$.
Under the assumption that the measurements are independent (which they might not be if, e.g., the device heated up during use), the likelihood has the form

$$\begin{aligned} P(\hat{A}_{1:N}|A_0, \sigma, \mathrm{N}, K) &= \prod_{i=1}^{N} P(\hat{A}_i|A_0, \sigma, \mathrm{N}, K) \\ &= \prod_{i=1}^{N} \mathrm{N}(\hat{A}_i; A_0, \sigma^2) \\ &= \prod_{i=1}^{N} \frac{1}{(2\pi)^{1/2}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{\hat{A}_i - A_0}{\sigma}\right)^2\right]. \end{aligned}$$

The resultant posterior is then

$$\begin{aligned} P(\sigma|\hat{A}_{1:N}, A_0, \mathrm{N}, K) &\propto \Theta(\sigma) \prod_{i=1}^{N} \frac{1}{(2\pi)^{1/2}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{\hat{A}_i - A_0}{\sigma}\right)^2\right] \\ &\propto \Theta(\sigma)\, \frac{1}{\sigma^N} \exp\left[-\frac{1}{2\,\sigma^2} \sum_{i=1}^{N} (\hat{A}_i - A_0)^2\right] \end{aligned}$$

[This question continues on the next page . . .]

$$= \Theta(\sigma)\,\frac{1}{\sigma^N}\exp\left(-\frac{1}{2}\frac{\hat{\sigma}^2}{\sigma^2/N}\right),$$

where the fact that $N \gg 1$ ensures that the posterior is proper and the summary statistic $\hat{\sigma}^2 = 1/N\sum_{i=1}^{N}(\hat{A}_i - A_i)^2$ has (optionally) been used to simply the final form.

INCLUDED FOR COMPLETENESS/REFERENCE ONLY:

The normalization constant can be obtained by integrating using a change of variables to $x = 1/\sigma^2$ (which follows a gamma distribution) to yield

$$P(\sigma|\hat{A}_{1:N}, A_0, \mathrm{N}, K) = \frac{2\,(N\hat{\sigma}^2/2)^{(N-1)/2}}{\Gamma[(N-1)/2]}\,\Theta(\sigma)\,\frac{1}{\sigma^N}\exp\left(-\frac{1}{2}\frac{\hat{\sigma}^2}{\sigma^2/N}\right),$$

where $\Gamma(x)$ is the standard gamma function.

[7 marks]

(b) What is the implied (predictive) probability distribution for some future measurement of this reference source, $P(\hat{A}|A_0, \hat{A}_{1:N}, \mathrm{N}, K)$? [This again can be calculated up to a normalization constant.]

ANSWER: (SIMILAR TO SEEN)

If σ were known then the predictive distribution would be simply

$$\begin{aligned} P(\hat{A}|A_0, \hat{A}_{1:N}, \sigma, \mathrm{N}, K) &= P(\hat{A}|A_0, \sigma, \mathrm{N}, K) \\ &= \mathrm{N}(\hat{A}; A_0, \sigma^2). \end{aligned}$$

However, the uncertainty about σ means it must be margnilised over, to give

$$P(\hat{A}|A_0, \hat{A}_{1:N}, \mathrm{N}, K)$$

$$= \int_0^\infty \mathrm{d}\sigma\, P(\sigma|\hat{A}_{1:N}, \mathrm{N}, K)\, P(\hat{A}|A_0, \sigma, \mathrm{N}, K)$$

$$= \int_0^\infty \mathrm{d}\sigma\, \frac{2\,(N\hat{\sigma}^2/2)^{(N-1)/2}}{\Gamma[(N-1)/2]}\,\frac{1}{\sigma^N}\exp\left(-\frac{1}{2}\frac{\hat{\sigma}^2}{\sigma^2/N}\right)\frac{1}{(2\pi)^{1/2}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{\hat{A}-A_0}{\sigma}\right)^2\right]$$

$$= \frac{2^{1/2}(N\hat{\sigma}^2/2)^{(N-1)/2}}{\pi^{1/2}\,\Gamma[(N-1)/2]}\int_0^\infty \mathrm{d}\sigma\,\frac{1}{\sigma^{N-1}}\exp\left[-\frac{1}{2}\frac{\hat{\sigma}^2 + (\hat{A}-A_0)^2}{\sigma^2/N}\right]$$

$$= \frac{1}{(\pi N)^{1/2}}\,\frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}\left[1 + \frac{(\hat{A}-A_0)^2}{N\hat{\sigma}^2}\right]^{-N/2}.$$

[7 marks]

(c) Is the assumption that the noise is normally distributed, N, justified on the basis of the available information? Explain your answer.

ANSWER: (SEEN)

[This question continues on the next page . . .]

On the basis of the information provided this assumption is not strongly justified. Arguments that might be invoked to support the assumption of normality, along with rebuttals, include:

- Maximum entropy arguments cannot be applied as there is particular evidence that $P(\hat{A}|A, D)$ has a defined mean or variance, let alone what values they take.
- The central limit theorem similarly cannot be invoked as there is no evidence that the device works by averagin/adding a large number of (independent) inputs.
- Even more spurious would be to argue that the normal distribution is very commonly used in statistics and/or that it has convenient algebraic properties.

[2 marks]

(d) This assumed model for $P(\hat{A}|A_0, D)$ implies that the device's outputs are unbiased. Is there evidence for this?

ANSWER: (SEEN)
No. There might be evidence from the set of test measurements, but these have not been numerically specified, and there is no prior knowledge about this; in particular, the statement about precision does not imply a lack of bias.

[1 mark]

(ii) The engineer notices that there are a few outliers in the data and so considers using a scaled Student-t distribution, with an additional shape parameter which controls how heavy the tails are.

(a) Why would a heavy-tailed distribution be appropriate in the presence of outliers?

ANSWER: (SEEN)
The main desirable property of a heavy-tailed distribution in the presence of outliers is that the core of the posterior distribution is not overly moved by a small number of outliers. An alternative way of looking at this is that if the measurement is an outlier the true value is not overwhelming ruled out.

[1 mark]

(b) Would a heavy-tailed distribution yield valid results if there were no outliers?

ANSWER: (SEEN)
Yes, although with a slightly decreased precision relative to the results which would be obtained under the normal model (for the same data).

[1 mark]

[This question continues on the next page . . .]

(c) Outline in point form how the engineer could use the test data to assess which of these two noise models is more appropriate, in particular accounting for the fact that both models have unspecified internal parameters.

ANSWER: (SIMILAR TO SEEN)
There are number of important ingredients in this calculation:

- Prior probabilities must be assigned to the two models; these could both reasonably be set to be 1/2, in which case this prior would cancel out in all subsequent calculations, but this cannot be considered automatic.
- Within each model prior distributions must be adopted for the unspecified internal parameters: location and scale for the normal; location, scale and shape for the Student-t. Given that both distributions are symmetric, using the same prior for the location parameter is well motivated. The fact that both have scale parameters suggests that the prior should be the same on these, although the interplay with the shape parameter in the case of the Student-t distribution makes that less clear. Particular care must be taken with the Student shape parameter as, for all sufficiently high values, the density is the same (and equal to a normal, although that is not so important).
- The marginal likelihoods for the data under the two models must be calculated.
- The two models can be compared using any of: the Bayes factors; the posterior odds; or one of the two posterior probabilities. It might then be appropriate to adopt whichever of the two models is favoured or operate with a weighted mixture of the two, but this model selection is a distinct second step beyond the comparison.

In this case the fact that the two models are nested (as the Student-t density tends to the normal density in the limit that the shape parameter goes to infinity) means that several other options are available too:

- If the priors on the location and scale parameters within the two models were the same, the Savage-Dickey density ratio could be used to perform the above calculation more efficiently.
- It would also be possible to recast this problem as a parameter inference problem within the Student model, with the location and scale parameters treated as nuisance quantities and the shape parameter being the focus.

[5 marks]

(iii) The engineer subsequently decides that both the above approaches are too restrictive, and instead decides to explore a "non-parametric" model in which $P(\hat{A}|A_0, D)$ is represented as a (potentially-infinite) mixture model of normal densities with a Dirichlet Process prior on their weights.

[This question continues on the next page . . .]

(a) Describe the mathematical structure of this model, highlighting the full set of parameters which should be included in the inference calculation.

ANSWER: (SEEN)
The basic structure of the model is hierarchical; the structure can be specified from the top down (e.g., by defininining a sampling scheme) or bottom up (e.g., in terms of distributions).
In terms of a sampling scheme the model could be defined by specifying sampling distributions for the number of kernels, M, the weights on these kernels, $W_{1:M}$, and then their means $\mu_{1:M}$ and variances, $\sigma^2_{1:M}$, which yields

$$\begin{aligned} M &\sim \pi_M \\ \alpha &\sim p_\alpha \\ W_{1:M} &\sim \text{DP}(M, \alpha) \\ \mu_m &\sim \pi_\sigma; m \in \{1, 2, \ldots, M\} \\ \sigma_m &\sim \pi_\mu; m \in \{1, 2, \ldots, M\}, \end{aligned}$$

where α is the Dirichlet Process parameter. There is considerable freedom in the choice of priors for the α, number of kernels, π_M, and their means π_σ and variances, π_μ; the key point here is the structure of the model rather than any specific distributions (except for the Dirichlet Process).
The full list of parameters which need to be inferred is $(M, W_{1:M}, \mu_{1:M}, \sigma_{1:M})$.

[3 marks]

(b) Explain how this model can be set up to avoid over-fitting of the data.

ANSWER: (UNSEEN)
This model certainly admits over-fitted densities, most simply if $M = N$ and a narrow kernel is associated with each of the data; the maximum likelihood model would be of this form with delta function kernels (i.e., $\sigma_m \to 0$).
However the prior on the variances (in particular, but also the number of kernels) should be chosen so that such models have a vanishingly small prior relative to the "just so" model described above.

[3 marks]

[Total 30 marks]

2. A high-school physics class is given an assignment to measure the speed of sound, s, and the speed of light, c. They do this by timing echoes (sound) and reflections (light) from the wall of a nearby building a distance of $D = 490$ metres from their classroom. The round trip travel time for a signal travelling at speed v (here equal to either s or c) would be $T = 2\,D/v$, but it is of course impossible to measure this time perfectly. The dominant source of uncertainty – the only one which needs to be considered here – is that the students are making their timings using an old-style digital watch which only displays seconds. A measured time of, say, $\hat{T} = 14$ would only provide the constraint that 13.5 seconds $\leq T <$ 14.5 seconds, implying that the $P(\hat{T}|T)$ is a uniform distribution.

(i) The students decide their only secure prior knowledge, K, is that light travels faster than sound (as, e.g., thunder is heard after lightning). They hence adopt an improper prior distribution of the form

$$P(s, c|K) \propto \Theta(s)\,\Theta(c)\,\Theta(c - s),$$

where $\Theta(x)$ is the Heaviside step function, equal to 0 if $x < 0$ and 1 if $x \geq 0$.

(a) Assess if this distribution effectively represents the stated prior knowledge and/or if it encodes additional information beyond that explicitly listed.

ANSWER: (SEEN)
The prior knowledge that light travels faster than sound is effectively encoded in the $\Theta(c - s)$ term.
The $\Theta(s)$ and $\Theta(c)$ terms additionally encode the fact that speeds are non-negative, which is reasonable but *not* stated explicitly.
Additionally, the $\Theta(c)$ term is redundant and could be omitted as the other two step functions between them enforce the fact that $c \geq 0$. [3 marks]

(b) Give two reasons why it can be useful to use an improper prior.

ANSWER: (SEEN)
Some reasons for using improper prior distributions are:

- They avoid the imposition of abitrary cut-offs (i.e., a maximum value for s or c here) that would represent conditioning on hypothetical assumed facts.
- They are mathematically convenient, allowing normalization terms to be omitted, resulting in simpler calculations.
- They enable ignorance (in this case there being no reason to prefer any valid value of s or c over any other valid value) even over an uncountable set to be encoded.

[2 marks]

[This question continues on the next page . . .]

(c) State two potential difficulties of using an improper prior.

ANSWER: (SEEN)
Some potential problems with improper priors are:

- They are subject to marginalization paradoxes (that are avoidable but can be easy to miss).
- They are not valid probability distributions as the probability that the quantity of interest is in the full set of values is infinite, whereas it should be unity.
- They can associate infinite weight with identical predictions (although this is not the case here) depending on the mapping from the chosen parametric representation.

[2 marks]

(ii) The students start by measuring the speed of sound, obtaining $\hat{T} = 3$.

(a) What is the implied posterior distribution, $P(s|\hat{T}, D, K)$?

ANSWER: (SIMILAR TO SEEN)
The posterior distribution is, by Bayes's theorem, proportional to

$$P(s|\hat{T}, D, K) \propto P(s|K)\, P(\hat{T}|s, D, K).$$

The relevant component of the above prior is $P(s|\hat{T}) \propto \Theta(s)$. The likelihood can be written in various different ways, some options for which are

$$\begin{aligned} P(\hat{T}|s, D, K) &= \mathrm{U}\left(\hat{T}; \frac{2D}{s} - \frac{1}{2}, \frac{2D}{s} + \frac{1}{2}\right) \\ &= \Theta\left[\hat{T} - \left(\frac{2D}{s} - \frac{1}{2}\right)\right] \Theta\left[\left(\frac{2D}{s} + \frac{1}{2}\right) - \hat{T}\right], \end{aligned}$$

where $\mathrm{U}(x; a, b)$ is a uniform distribution with $a \leq b$.
Combining these two components and discarding irrelevant terms yields

$$\begin{aligned} P(s|\hat{T}, D, K) &\propto P(s|K)\, P(\hat{T}|s, D, K) \\ &\propto \Theta(s)\, \Theta\left[\hat{T} - \left(\frac{2D}{s} - \frac{1}{2}\right)\right] \Theta\left[\left(\frac{2D}{s} + \frac{1}{2}\right) - \hat{T}\right] \\ &= \Theta\left(s - \frac{2D}{\hat{T} + 1/2}\right) \Theta\left(\frac{2D}{\hat{T} - 1/2} - s\right). \end{aligned}$$

This posterior distribution can then be normalized to give

$$\begin{aligned} P(s|\hat{T}, D, K) &= \frac{1}{2D\,[1/(\hat{T} - 1/2) - 1/(\hat{T} + 1/2)]} \Theta\left(s - \frac{2D}{\hat{T} + 1/2}\right) \Theta\left(\frac{2D}{\hat{T} - 1/2} - s\right) \\ &= \mathrm{U}\left(s; \frac{2D}{\hat{T} + 1/2}, \frac{2D}{\hat{T} - 1/2}\right) \\ &= \mathrm{U}\left(s; 280.0\,\mathrm{m/s}, 392.0\,\mathrm{m/s}\right), \end{aligned}$$

[This question continues on the next page . . .]

where the final expression uses the supplied numerical values for $\hat{T}$ and D. [A less formal route to the final result would be acceptable, but it cannot simply be stated – it is important to demonstrate how the data drive the final constraints on s.] [6 marks]

(b) Does the prior distribution cause any problems? Explain your answer.

ANSWER: (SIMILAR TO SEEN)
The fact that the prior in s is improper is largely resolved by the tightly constraining likelihood, resulting in a proper posterior.
However, if c were included in the calculation by finding the joint posterior and then attempting to marginalize, the result would be a "non-conglomerability" paradox. One way to avoid this is to simply not to consider c at all in this calculation; another option would be to adopt a maximum prior value for the speed of light $c_{\max}$ and then, after doing the full calculation (including marginalization), to take the limit $c_{\max} \to \infty$.

[2 marks]

(iii) The students now turn to measuring the speed of light but the reflection appears so rapidly that they measure $\hat{T} = 0$.

(a) What is the implied posterior distribution, $P(c|\hat{T}, D, K)$?

ANSWER: (UNSEEN)
The calculation of the posterior is structurally similar to that above, although with some important distinctions, so it given in full here.
The posterior distribution is, by Bayes's theorem, proportional to

$$P(c|\hat{T}, D, K) \propto P(c|K)\, P(\hat{T}|c, D, K).$$

The relevant component of the above prior is $P(c|\hat{T}) \propto \Theta(c)$. The likelihood can, as before, be written variously as

$$\begin{aligned} P(\hat{T}|c, D, K) &= \mathrm{U}\left(\hat{T}; \frac{2D}{c} - \frac{1}{2}, \frac{2D}{c} + \frac{1}{2}\right) \\ &= \Theta\left[\hat{T} - \left(\frac{2D}{c} - \frac{1}{2}\right)\right] \Theta\left[\left(\frac{2D}{c} + \frac{1}{2}\right) - \hat{T}\right]. \end{aligned}$$

Combining these two components for the specific case that $\hat{T} = 0$ then yields

$$\begin{aligned} P(c|\hat{T} = 0, D, K) &\propto P(c|K)\, P(\hat{T} = 0|c, D, K) \\ &\propto \Theta(c)\, \Theta\left[\hat{T} - \left(\frac{2D}{c} - \frac{1}{2}\right)\right] \Theta\left[\left(\frac{2D}{c} + \frac{1}{2}\right) - \hat{T}\right] \\ &\propto \Theta(c)\, \Theta\left[-\left(\frac{2D}{c} - \frac{1}{2}\right)\right] \Theta\left(\frac{2D}{c} + \frac{1}{2}\right) \end{aligned}$$

[This question continues on the next page . . .]

$$
\begin{aligned}
&= \Theta\left(c - \frac{2\,D}{1/2}\right) \times 1 \\
&= \Theta(c - 4\,D) \\
&= \Theta(c - 1690\,\mathrm{m/s}),
\end{aligned}
$$

where the final expression uses the numerical value of the distance. The result that the measurement of $\hat{T} = 0$ yields a lower limit on the speed of light of 1690 m/s is sensible; but the result is not a proper probability distribution as neither the prior knowledge nor the data provide an *upper* limit on c.

[4 marks]

(b) Does the prior distribution cause any problems? Explain your answer.

ANSWER: (SIMILAR TO SEEN)
The use of an improper prior here is problematic, as already seen in mathematical terms from the previous answer. More qualitatively, both the prior and the likelihood only encode lower limits on c, but neither provide either a hard any upper limit (nor, in combination, sufficient constraints that the upper tail of the posterior distribution decreases sufficiently rapidly with D to be normalizeable). [3 marks]

(c) If the students made repeat measurements and they all yielded $\hat{T} = 0$, would the implied constraints on c be improved? Explain your answer.

ANSWER: (UNSEEN)
No: the step-like nature of the likelihood means that there would be no narrowing of the posterior with sample size. [1 mark]

(iv) One of the students re-measures the distance to the building and gets a different result from that quoted above, implying some potentially significant uncertainty about the value of D. Outline in point form how this additional source of uncertainty could be properly and robustly incorporated into the data analysis.

[Do *not* attempt a full calculation.]

ANSWER: (SIMILAR TO SEEN)
The unknown value of D should be incorporated as a nuisance parameter and then marginalized out, which can be done by:

1. Adopt a (presumably) uninformative prior distribution for D, such as $P(D|K) = \Theta(D)$. Any distribution which was i) smooth and ii) broad enough that the two measurements are plausible would be reasonable; but to ensure the final result is robust it would be important to repeat the calculation with different forms for this prior (or to include a prior on the space of these distributions, although that would be overkill).

[This question continues on the next page . . .]

2. Assume a distribution for $P(\hat{D}_1, \hat{D}_2|D, K)$ that describes the effective likelihood for the old and new distance measurements.

s would likely be unchanged.

3. Include D in the inference by calculating the joint posterior distribution $P(s, c, D|\hat{T}, \hat{D}_1, \hat{D}_2, K)$.
4. Marginalize over the nuisance parameter D by integrating over it to obtain

$$P(s, c|\hat{T}, \hat{D}_1, \hat{D}_2, K) = \int_0^\infty \mathrm{d}D\, P(s, c, D|\hat{T}, \hat{D}_1, \hat{D}_2, K).$$

5. Optionally marginalize over either s or c, although this would likely run into some of the problems identified above.

[7 marks]

[Total 30 marks]

MATH70090 Bayesian Methods

Question	Marker's comment
1	This question was broadly answered well, although with the calculations done systematically better than the more conceptual questions.Part i a: This was done well (as expected given that this is a standard parameter inference calculation).Part i b: Most students correctly formulated this as an integral marginalising over the unknown value of sigma; but the subsequent calculation was not done so well.Part i c: Most students saw that there was nothing in the information provided to justify this in terms of whatever process induces the noise; some invoked maximum entropy principles, which are not appropriate in this case.Part i d: Most students saw there was no strong evidence for the noise being unbiased.Part ii a: This was answered well.Part ii b: This was answered well.Part ii c: This was answered well, with the exception of the need to adopt prior distributions for the Student-t shape/d.o.f. parameter.Part iii a: This was answered well (unsurprisingly, given that this was effectively contained in the lecture notes which the students could bring to the exam).Part iii b: This sub-question, the most challenging part of the problem, was not answered well.
2	This question was answered reasonably well, although the unusual likelihood (which had been seen in lectures/problems) did cause some confusion.Part i a: Answered well, although a few students did not see that the Theta(s) and Theta(c) terms encoded that the speeds are non-negative - reasonable, but not explicitly listed.Part i b: Answerednbsp;well.Part i c: Answered reasonably well.Part nbsp;ii a: Answered well.Part ii b: The answers to this subtle and challenging question were somewhat confused, covering a broad range of views - the basic issue is that the constraint on s is clear, but it cannot be obtained by marginalising the joint distribution in s and c over c.Part iii a: One of the more challenging parts of this question, which was not answered so well, with few students correctly manipulating the step functions to obtain a lower bound on c.Part iii b: Most students correctly noted that the improper prior yields an improper posterior in this case as the data provide only a lower limit on c.Part iii c: This was mostly answered well, although some students did not see that the fact that the reported time is not a stochastic function of the actual time means that repeated observations in this case provide no extra information.Part iv: This was broadly answered well, but few students gave all the critical ingredients, which include the addition of D to the parameter set, some model for the sampling distributions of the measurements/estimates of D, and the need to marginalise over its unknown value.