UNIVERSITY OF CAMBRIDGE

MPhil in Data Intensive Science

Lent Term 2024

S2 Major Module Coursework
Christopher J. Moore

This problem is based on the published papers referenced below. You are expected to use the public data release associated with Reference 1 available at https://dataverse.harvard.edu/dataverse/amelc.

*Attempt **all** parts of the coursework.*

The anticipated number of marks allocated to each part of a question is indicated in the right margin, to be assessed from the code you provide and your report.

*This coursework should be submitted via a GitLab repository which will be created for you. Place all of your code and your report into this repository. The report should be in pdf format and placed in a folder named `report`. You will be provided access to your repository until the **deadline of 23:59, Friday 4th April, 2025**, after which your work will be deemed ready to assess.*

You are expected to submit code and associated material that demonstrates good software development practices as covered in the Research Computing module.

Your report should not exceed 3000 words (including tables, figure captions and appendices but excluding references); please indicate its word count on its front cover. You are reminded to comply with the requirements given in the Course Handbook regarding the use of, and declaration of use of, autogeneration tools.

The Antikythera mechanism is an ancient (c. 100 BC) mechanical calculator, excavated from a shipwreck in the Mediterranean in 1901, believed to have been used for predicting the motions of the Sun, Moon and planets among other astronomical phenomena. The original purpose and functionality of the mechanism are not fully understood. However, it is possible to infer some details from the incomplete parts of the mechanism that survive.

The Antikythera mechanism included a calendar ring with holes punched around its circumference; see figure 1. About 25% of the ring survives, albeit fractured into several sections that have shifted relative to each other and become misaligned. It was previously thought that the complete ring contained 365 holes and functioned as a solar calendar. However, it has recently been suggested that it was a lunar calendar with 354 holes, see Ref [1].

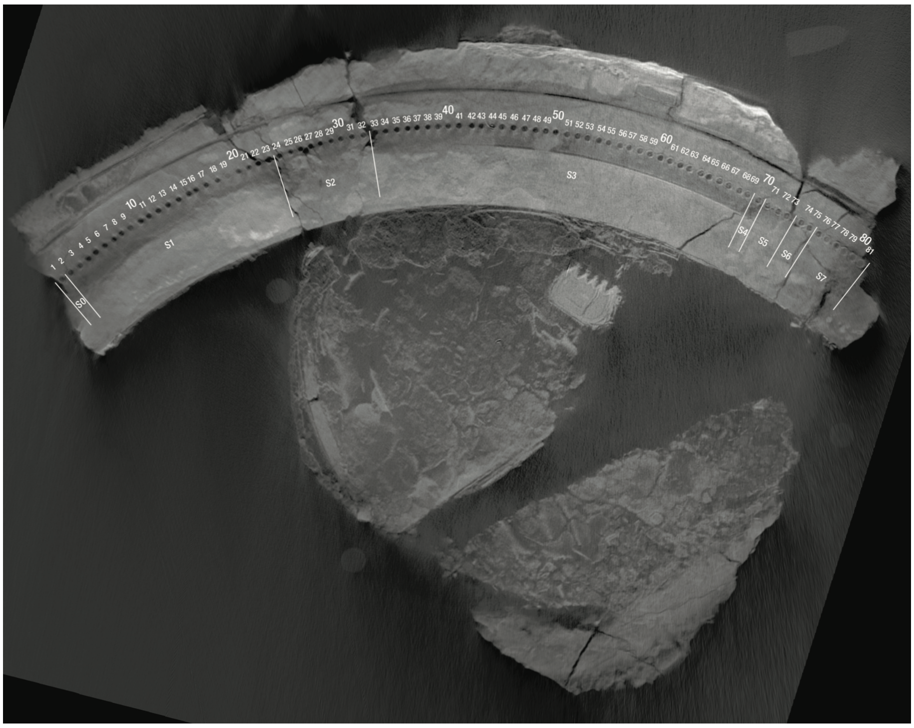

Figure 1: The calendar ring. Figure reproduced from Budiselic *et al.* (2020).

1 The objective of this coursework project is to use an X-ray image of the calendar ring and Bayesian inference with Hamiltonian Monte Carlo to infer the number of holes that were in the complete calendar ring, thereby learning something about its intended purpose. As a starting point for this analysis [parts (a) to (c) of this question] we will be following the Bayesian reanalysis of the X-ray image data in Ref [2].

(a) Download the data from Ref [3]: two images, a data file with the measured hole locations, and a README file which you should read carefully. Make a plot showing the measured hole locations $d_i \in \mathbb{R}^2$ in the x-y plane, where i labels the holes, indicating clearly which holes are in each fractured section of the ring. [4]

(b) You should use the model for the hole locations described in section 2 of Ref [2]. For simplicity, it is assumed that all parts of the disk lie in the x-y plane. This model assumes that there were originally N holes, arranged regularly around a circle of radius r. The broken sections of the ring are allowed to be misaligned by translations and rotations in the plane. [5]

[*The rest of this problem focuses on the use of MCMC and gradient-based methods. Therefore, it will be easiest if you treat N as a continuous parameter. One way to think about this is allowing for the possibility that the hole spacings have a single discontinuity at one point in the ring.*]

(c) You should use a Gaussian likelihood function for each hole location and assume that the error in the placement of different holes is independent.

$$\mathcal{L}(D|\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{2}[\boldsymbol{d}_i - \boldsymbol{m}_i]^{\mathrm{T}} \cdot \boldsymbol{\Sigma}^{-1} \cdot [\boldsymbol{d}_i - \boldsymbol{m}_i]\right)}{2\pi|\boldsymbol{\Sigma}|}$$

The data is the set of all measured hole locations, $D = \{\boldsymbol{d}_i | i = 1, \ldots, N_{\text{holes}}\}$. The number of holes in the surviving part of the ring is $N_{\text{holes}} = 81$. The product is over the measured holes and $\boldsymbol{m}_i$ denotes the model prediction (a function of $\boldsymbol{\theta}$) for the location of hole i.

For the covariance matrix $\boldsymbol{\Sigma}$ you should consider two options. Firstly, you should consider a 2-dimensional isotropic Gaussian distribution with standard deviation σ. Secondly, you should consider a 2-dimensional Gaussian distribution with principle axes aligned with the radial and tangential directions for each hole and with standard deviations σ_r and σ_t describing the radial and tangential accuracies of the hole placement. This model is described in section 2 of Ref [2].

Overall, the model includes several parameters: the ring radius r, the number of holes in the original ring N, one or two σ-parameters describing the covariance matrix $\boldsymbol{\Sigma}$, and 3 parameters (x_j, y_j, α_j) for each section of the ring. Here, the index j labels labels the fragmented sections of the ring, $j \in \{0, 1, 2, \ldots, 7\}$. The collection of all these model parameters is denoted $\boldsymbol{\theta}$. [5]

(d) In addition to the log-likelihood, your code should be able to make use of its gradient,

$$\frac{\partial}{\partial \theta^{\mu}} \log \mathcal{L}(D|\boldsymbol{\theta}),$$

where θ^{μ} denotes the components of the parameter vector $\boldsymbol{\theta}$. Your report should included details of how the derivatives were calculated and brief details of any code tests performed. [10]

(TURN OVER

This model and likelihood functions are sufficiently simple that it is possible to evaluate all of the derivatives analytically and to code them up as separate functions. Alternatively, you can make use of automatic differentiation, e.g. using `jax`*. Either approach can be used.*

(e) Find the maximum likelihood parameters for both models. (That is, the model with isotropic covariance matrix and the model with radial/tangential covariance matrix). On your plot of the measured hole locations (or else on another copy of this plot, if overcrowding becomes an issue), add new markers showing the maximum likelihood predictions from both models for the hole locations. [8]

(f) Sample the posterior distribution for both models using any variation of the Hamiltonian Monte Carlo sampling algorithm. Pick one of the holes, on your plot of the measured hole locations, show the posterior predictive distribution for its hole locations using both models. [8]

(g) Explain the role of the 2-dimensional covariance matrix $\boldsymbol{\Sigma}$. Explain why the radial/tangential model for the covariance matrix might *a priori* be considered a good model. Does the data favour this radial/tangential model over the simple isotropic model? Your answer should be justified using appropriate plots and/or quantitative measurements. [10]

References

The original paper, for general background information.
[1] C. Budiselic, A. T. Thoeni, M. Dubno & A. T. Ramsey, "Antikythera Mechanism Shows Evidence of Lunar Calendar", The Horological Journal, March 2021, https://doi.org/10.31235/osf.io/fzp8u

The Bayesian reanalysis of the problem.
[2] G. Woan & J. Bayley, "An Improved Calendar Ring Hole-Count for the Antikythera Mechanism: A Fresh Analysis", The Horological Journal, July 2024, https://arxiv.org/abs/2403.00040

The public data.
[3] A. Thoeni, C. Budiselic & A. T. Ramsey, "Replication Data for: Antikythera Mechanism Shows Evidence of Lunar Calendar", version 3, 2019, https://doi.org/10.7910/DVN/VJGLVS