Probabilistic Reasoning

Assignment C, 2022

Practical information

The assignments for this course are graded as INSUFFICIENT, SUFFICIENT, GOOD, or EXCELLENT. The course website (tab 'Examination and Grading') describes how these results contribute towards your final course grade.

The assignments are all about demonstrating understanding; explanation of your answers is therefore far more important than the answers themselves. Without explanations you will receive no more than an insufficient grade.

Warning: The different assignments in this course require an increasing amount of work, so you are advised to partially work on them in parallel instead of starting to work on one only after the deadline of another. The assignments not depend on each other, so this is feasible.

▷ Goal

The goal of Assignment C is to give you experience using the probabilistic programming language Stan to solve a practically motivated Bayesian data analysis problem.

▷ Requisites

- The assignments **must** be done in groups of 2 students.
- The **deadlines** are given on the group website.
- Submissions should be uploaded as a zip-file on Blackboard, containing
  - *the core submission:* **a single pdf report** with all text answers, tables/visualisations, and requested code snippets (Stan, R/Python) inlined, organised by numbered question; this pdf should contain all the information needed by a grader to evaluate your assignment;
  - *the supporting code:* **a folder with the Stan and R/Python code** that you used to perform your analyses, clearly annotated with the

relevant subquestions using code comments; this code serves the function to make your results reproducible in the spirit of reproducible science.

▷ Grading

This assignment is graded INSUFFICIENT, SUFFICIENT, GOOD, or EXCELLENT. Each subquestion (e.g. 2(b)) is worth 2 points. 18 points equals an EXCELLENT. As such, question 5(a) can be viewed as being worth 2 bonus points.

▷ Software

- In this assignment, you will be writing statistical models in Stan (see https://mc-stan.org/). Further, you will need to call Stan from a scripting language (recommendation: R or Python) to feed it data and to postprocess and graph its results (see https://mc-stan.org/users/documentation/).
- In your analyses in Stan, please make sure to check the Gelman-Rubin statistic $\hat{R}$ and effective sample size N_{eff} to know when not to trust your results. You are not expected to perform posterior predictive checks or SBC.
- When using Stan, you might want to use the documentation listed here: https://mc-stan.org/users/documentation/ . Particularly the Case Studies and User's Guide might be helpful.

▷ Data

- You can find the data you need for the analyses in this assignment on Blackboard. They are a compilation of experimental results into the $\delta^{18}O$-temperature relationship for carbonates formed by various organisms (see Background section below for some context). You might notice that the data file contains more variables than you need in this assignment. Please feel free to ignore any variables you think you do not need to answer the questions.

▷ Background - a practical Bayesian data analysis problem

Climate reconstructions of the geological past

Anthropogenic carbon emissions and land use change have resulted in rapid increases in atmospheric CO_2 levels and a global mean temperature increase of 1.1°C (as of 2021) since the Industrial Revolution (1850 CE), issuing a global

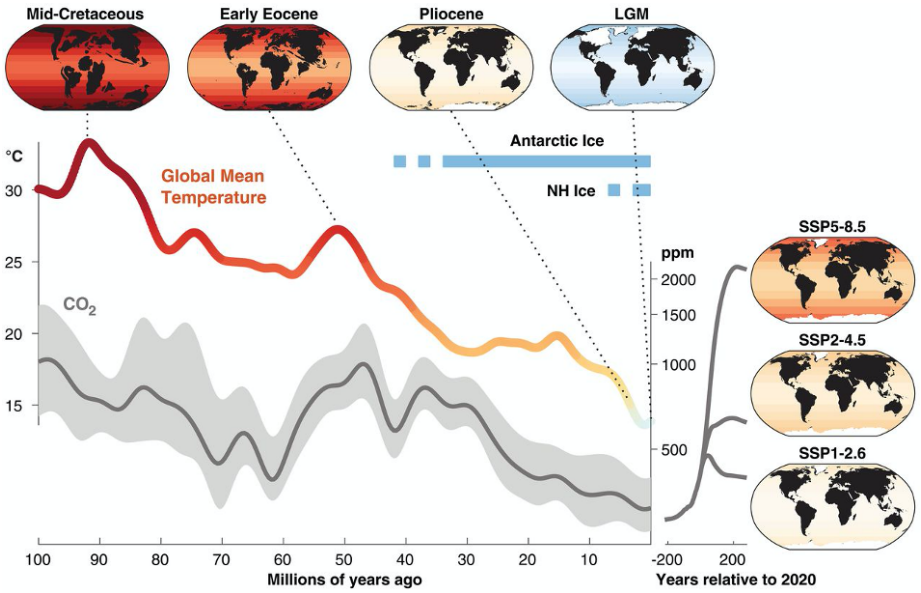

Figure 1: Overview of atmospheric CO_2 concentrations and global mean temperature during the last 100 million years compared to projected CO_2 concentrations under varying future emission scenarios (modified after Tierney et al., 2020).

climate crisis (IPCC, 2021). Unless global net-zero carbon emissions are reached in the coming decades, climate models project global mean temperature to further increase to 2°C - 5°C above pre-industrial levels, reaching a climate state unprecedented in the last tens of millions of years (Tierney et al., 2020; see Figure 1). Reconstructions of climate conditions during past geological periods with similar atmospheric CO_2 levels are indispensable for understanding the impact of this greenhouse warming on our climate, improving models for future climate projections, and informing policymakers about smart legislation needed to prevent and mitigate the effects of catastrophic climate change.

Oxygen isotopes, a intertemporal thermometer

Reconstructions of sea water temperatures in Earth's geological past (millions of years back in time) largely rely on the analyses of the oxygen isotope composition (ratio of ^{18}O to ^{16}O) in the fossil carbonate skeletons of marine organisms ranging from tiny unicellular algae producing microscopic skeletons out of calcium carbonate ($CaCO_3$) to large (meter-scale) corals and mollusk shells (e.g. Pearson et al., 2001; de Winter et al., 2020). The isotopic composition of these skeletons ($\delta^{18}O_c$) is known to depend (via a function f) on two factors:

1. the isotopic composition of the sea water in which the organism grows

($\delta^{18}O_w$) and

2. the temperature (T) at which the carbonate is formed:

$$\delta^{18}O_c = f(\delta^{18}O_w, T),$$

where $\delta^{18}O_x = \frac{([^{18}O]/[^{16}O])_x}{([^{18}O]/[^{16}O])_{reference}}$ is the ratio of oxygen isotopes in the sea water ($x = w$) or the carbonate ($x = c$) divided by the oxygen isotope ratio in some reference sample.

The oxygen isotope composition of the sea water is largely assumed to be known (or remain constant) through geological time, meaning that $\delta^{18}O_c$ analyses in fossil carbonate can be used as a proxy for temperatures in the geological past.

A statistical challenge

Unfortunately, use of the $\delta^{18}O_c$ proxy suffers from a calibration problem: Many previous studies have attempted to calibrate the proxy by analyzing carbonates formed at a known temperature, only to find different temperature relationships for carbonates produced by different organisms (or precipitated inorganically in a lab; see e.g. Kim and O'Neil, 1997; Marchitto et al., 2014) and for calcium carbonates with a different mineral structure (e.g aragonite vs. calcite; see Kim et al., 2007). On top of that, paleoclimatologists use these various relationships interchangeably and without propagating the calibration uncertainty, leading to irreconcilable temperature reconstructions which blur the relationship between atmospheric CO_2 concentrations and global warming throughout geological time.

In this assignment, we explore the $\delta^{18}O_c$ proxy and its relationship with the formation temperature of carbonates through a Bayesian modelling approach using the probabilistic programming language Stan, using regression models (such as hierarchical linear models).

The practical motivation for such modelling is that its results are used as an important input to reconstructions of Earth's temperature, such as the reconstruction over the past 66 million years (since the extinction of the non-avian dinosaurs) from the latest oxygen isotope compilation by Westerhold et al. (2020; see Figure 2), which are crucial for our understanding of Earth's temperatures in the past.

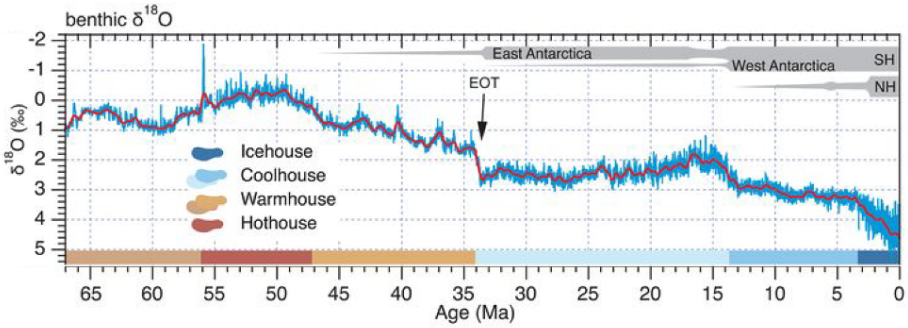

Figure 2: Compilation of oxygen isotope compositions ($\delta^{18}O$) of marine carbonates formed in the last 66 million years (Ma = million years ago). Lower oxygen isotope compositions correspond to warmer to ocean water temperatures (note the inverted vertical axis), which are an indicator of the global climate state. Grey bars on the top indicate the presence of polar ice sheets, based on geological information. The colors on the bottom of the plot highlight various climate states (modified after Westerhold et al., 2020).

Probabilistic Reasoning

Assignment C: probabilistic programming

Question form

1 Basic linear regression

We want to fit a basic linear regression to the data. That is, denoting the individual measurements with an index i between 1 and I,

$$T^{(i)} = a + b \cdot (\delta^{18}O_c{}^{(i)} - \delta^{18}O_w^{(i)}) + \sigma_T \cdot \epsilon^{(i)},$$

where $\epsilon^{(i)}$ are i.i.d. draws from a standard normal distribution ($normal(0, 1)$). (This functional form is predicted by chemical considerations.) In this assignment, you may treat σ_T as an unknown parameter, to be determined by the analysis, and you may treat the $\delta^{18}O_c{}^{(i)}$ and $\delta^{18}O_w{}^{(i)}$ measurements as precise, without any measurement error[1].

(a) Please write a Stan model to implement this linear regression and give a table of the posterior means of the estimated parameters a (the intercept), b (the slope) and σ_T (the residual uncertainty on T) as well as measures of the uncertainty on the parameters (posterior standard deviations).

2 Adding priors

We have prior knowledge, even before running these experiments, that marine temperatures T tend to lie (well) within the range of -2 to 50 degrees Celcius and that $\delta^{18}O_w - \delta^{18}O_c$ values tend to be in the range of -3 to 5. Finally, we can assume that we should be able to measure water temperatures at least within 2 degrees certainty.

(a) Using this knowledge, please specify some priors on the parameters of our regression model and validate your chosen priors using a prior predictive check.

(b) Do the priors meaningfully change the results of your regression of question 1?

3 Adding hierarchical structure

As you might have noticed, so far, our model has very limited predictive performance: if we measure $\delta^{18}O_c$ and $\delta^{18}O_w$ values, we have enormous residual uncertainty on the temperature T. As such, the model is not very useful for use

[1] Naturally, a more sophisticated analysis would revisit these assumptions. (See the outro.)

in temperature reconstructions. The problem is that the relationship between $\delta^{18}O_c$, $\delta^{18}O_w$ and T is highly dependent on the particular species j (where j lies between 1 and J) of organism we are looking at.

We can try to fix the issue by fitting a separate linear regression for each species j:

$$T^{(i)} = a_{j_i} + b_{j_i} \cdot (\delta^{18}O_c{}^{(i)} - \delta^{18}O_w^{(i)}) + \sigma_T \cdot \epsilon^{(i)},$$

where j_i is the species of measurement i and we allow for different intercepts a_j and slopes b_j for each species j. (You may assume that the σ_T does not depend on the species.)

(a) Please try this. What do you observe?

A better approach compared to the *complete pooling* of question 1 and the *no pooling* of question 2(a), is to *partially pool* measurements from different species together. We can achieve this by adding hierarchical structure to our regression model:

- assume that all the a_j are drawn from a $normal(a, \sigma_a)$-prior and all b_j from a $normal(b, \sigma_b)$-prior, where a, σ_a, b and σ_b are new parameters;
- put your prior from question 2 on a and b and choose a reasonable (weak) new prior for σ_a and σ_b (perhaps weakly informed by what your observed in the outcomes for a_j and b_j of the no-pooling model - no need to do another prior predictive check).

(b) Please implement this. Hint: you may want to use a non-centered parameterisation (https://mc-stan.org/docs/stan-users-guide/reparameterization.html #non-centered-parameterization) to aid the model to fit faster.

(c) How do the results change compared to the no pooling approach we followed earlier in this question? How do they compare to your complete pooling approach of question 1? Finally, according to your partial pooling analysis, do the values of the intercepts a_j meaningfully depend on the choice of species j, and how about the slopes b_j?

(d) Please explain in your own words: what are the (intuitive) interpretations of all the parameters a_j, a, σ_a (similarly, b_j, b, σ_b) and σ_T?

4 Using the model for temperature reconstruction

Suppose that we are geologists who have measured a sequence of new $\delta^{18}O_c{}^{(k)}$ and $\delta^{18}O_w^{(k)}$ values, for k between 1 and K, for some fixed species j (again, assuming no uncertainty on these measurements).

(a) Please implement a piece of code (in R, Python, or Stan itself, depending on your preference) that predicts, using the posterior of your hierarchical regression model, the corresponding temperatures $T^{(k)}$ together with the posterior-predictive uncertainty in these predicted temperatures. Please explain your code.

(b) Please alter your code (and explain it!) to account for uncertainty $\sigma_{\delta^{18}O_c}{}^{(k)}$ and $\sigma^{(k)}_{\delta^{18}O_w}$ in the $\delta^{18}O_c{}^{(k)}$ and $\delta^{18}O_w^{(k)}$ measurements - you can continue to assume that the $\delta^{18}O_c{}^{(i)}$ and $\delta^{18}O_w^{(i)}$ measurements used to fit your original regression model did not have any meaningful uncertainty associated with them.

5 Bonus: visualising the posterior

(a) Please visualise the posterior of the hierarchical model, by plotting a regression line for each species j, together with a visualisation of the uncertainty on the regression line that is implied by the posterior uncertainty on a_j and b_j.

Some final words (outro - not a question)

In this assignment, you have seen how Bayesian data analysis naturally progresses by iteratively building a model, criticising it, and refining it further in the light of the criticism. Natural next steps are to incorporate uncertainty on the $\delta^{18}O_c{}^{(i)}$ and $\delta^{18}O_w^{(i)}$ measurements in the model, to allow for correlation between a_j and b_j, to treat σ_T as data whenever it is specified in a study, to deal properly with missing data (such as unspecified measurement errors), and to add further levels to the hierarchy. Following this route leads to state-of-the-art models for performing the sort of climate reconstructions that are important inputs to e.g. IPCC reports.

References

1. de Winter, N. J. et al. The giant marine gastropod Campanile giganteum (Lamarck, 1804) as a high-resolution archive of seasonality in the Eocene greenhouse world. Geochemistry, Geophysics, Geosystems 21, e2019GC008794 (2020).

2. Kim, S.-T. O'Neil, J. R. Equilibrium and nonequilibrium oxygen isotope effects in synthetic carbonates. Geochimica et Cosmochimica Acta 61, 3461–3475 (1997).

3. Kim, S.-T., O'Neil, J. R., Hillaire-Marcel, C. Mucci, A. Oxygen isotope fractionation between synthetic aragonite and water: Influence of temperature and Mg2+ concentration. Geochimica et Cosmochimica Acta 71, 4704–4715 (2007).

4. Marchitto, T. M. et al. Improved oxygen isotope temperature calibrations for cosmopolitan benthic foraminifera. Geochimica et Cosmochimica Acta 130, 1–11 (2014).

5. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. (Cambridge University Press, 2021).

6. Pearson, P. N. et al. Warm tropical sea surface temperatures in the Late Cretaceous and Eocene epochs. Nature 413, 481–487 (2001).

7. Tierney, J. E. et al. Past climates inform our future. Science 370, (2020).