A Note on Paleoclimate Reconstruction, Ice Cores, and Probabilistic Modeling

Will Tebbutt

May 14, 2018

Abstract

This note introduces the problem of paleoclimate reconstruction, focusing on ice cores. It was written primarily for the authors own benefit, but it might be mildly interesting to anyone who knows something about probabilistic modeling but not paleoclimate reconstruction. I proceed to review some interesting work, highlight some relevant sources of data, and to present a Gaussian process model in what is essentially a toy problem. Note that this is very much a work-in-progress (WIP), so it will be updated regularly.

1 Introduction and Problem Description

The problem of Paleoclimate Reconstruction is that of using proxy data extending over a long period of history to infer the properties of past climate, such as temperature and green house gas (GHG) concentrations. For example, the ratio of certain isotopes of water trapped in ice cores (primarily) drilled in Antarctica and Greenland can be used to infer local historical temperature, and we can directly measure GHG concentrations. Other proxy measurements include tree rings , bore hole temperatures , and pollen records . Throughout this note we shall restrict our attention to ice cores.

get a good reference

get a good reference

get a good reference

BAS [2014] provides an excellent overview of the use of ice cores for Paleoclimate reconstruction, which I would highly recommend reading. Roughly speaking, to use an ice core to infer properties of past climate we must first map the depth at which a measurement of a proxy is made (i.e. how far down the core we make a measurement) to a point time, and subsequently determine a function relating whatever quantity is measured to whatever quantity we actually care about. For example, this is simply the identity function if we measure CO2 and we want to know about historical CO2 concentrations in the vicinity of the core location, but is not so simple for quantities that we cannot measure directly from the core, such as temperature.

Figure 1 (due to Smerdon and Kaplan [2007]) provides a depiction of a very general strategy for going about inferring the latter relationship. We first assume that the relationship between the proxy and our quantities of interest is temporally invariant, infer the relationship between our quantities of interest using contemporary proxy and instrumental data, and then use this inferred relationship to achieve our stated goal of inferring properties of the paleoclimate.

For an excellent overview of possible applications of machine learning in climate science, see Banerjee and Monteleoni [2014]. In particular see the "Paleoclimate" section, which provided the starting point for this note.

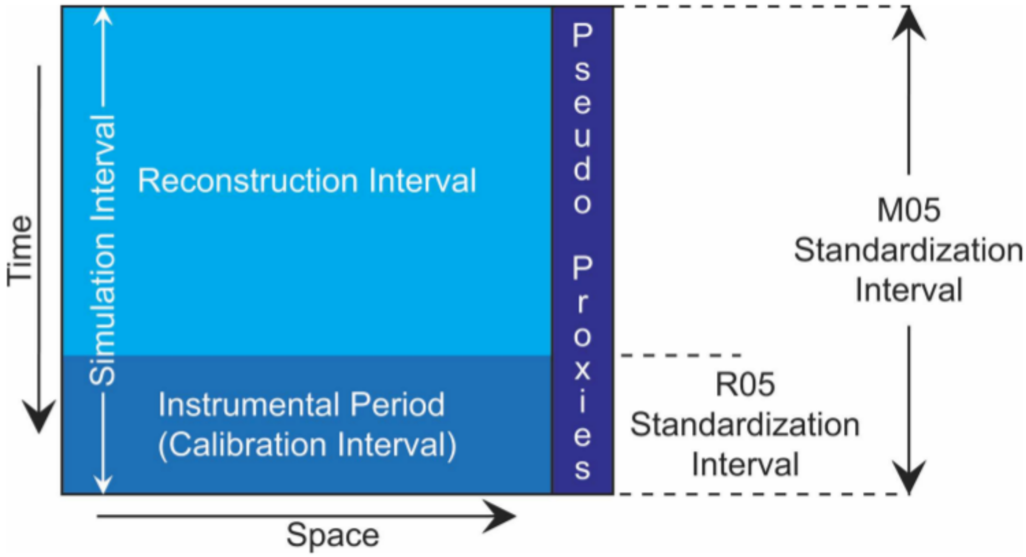

Figure 1: Credit: Smerdon and Kaplan [2007]. For the purposes of this note, both of the "Standardization Interval"s depictions can be safely ignored. We have access to instrumental data from the "Calibration Interval", and proxy data for the entire time series. We wish to construct a model using this data to infer the instrumental data, and derived quantities, for the "Reconstruction Interval". It is common practice to use simulated data to test the efficacy of such schemes, as we can construct train / test data for the entire "Simulation Interval". In such a scenario we must generate "Pseudo-Proxy" data.

2 Related Work

This section is *currently* an assortment of overviews of various papers, as opposed to a cohesive review that provides some amount of insight. May contain opinions.

Li et al. [2010] propose a relatively simple probabilistic model that elegantly integrates multiple different types of proxy data (bore holes, tree rings, and pollen records) and multiple drivers of temperature (GHG concentrations, volcanic activity, and solar irradiance) to reconstruct historical temperature. The model is validated on synthetic data from an atmosphere-ocean general circulation model (AOGCM). This paper is of interest as it seeks to integrate multiple sources of proxy data, the utility of which was apparently not clear / validated previously.

McShane and Wyner [2011] makes for interesting reading. A key contribution is the observation that, if we constrain ourselves to consider only linear + Gaussian models for paleoclimate reconstruction, a signal which is designed to have temporally short-term structure resembling that of the actual temperature record but *which is generated independently of the temperature record*, thus lacking the same long-term structure, has predictive power quite indistinguishable from any of the proxy records. Performing a more sophisticated analysis in which we perform Bayesian linear regression with an autoregressive term to handle correlated residuals yields substantially better results, but still struggles to capture the recent up-tick in NH surface temperature. The authors state that their key contribution is "to seriously grapple with the uncertainty involved in paleoclimatological reconstructions"; although I agree that they have broadly achieved this within the context of linear Gaussian models, I find the use of PCA (which they point out is standard practice) for dimensionality reduction unjustifiable as a pre-processing step as the strong i.i.d. assumptions it makes regarding the data generating process are quite clearly not applicable in this context. Consequently, I wonder whether the use of a different methodology for dimensionality reduction might not yield better results (i.e. include a dimensionality reduction step as part of the model). Moreover, it is not clear that linearity and Gaussianity are valid modeling assumptions here; they should be tested empirically. None of the above commentary should be taken to detract from what I feel is an extremely useful piece of work, that I would strongly recommend reading.

Tingley and Huybers [2013] is interesting Bayesian modeling of temperature. Again it's linear and Gaussian though.

Doan [2015] is a thesis containing quite a lot of ground-work for the integration of multiple paleoclimatic time series that you might find in the wild. The individual papers are discussed below, but on its own it provides an interesting reference regarding inference in the GP corresponding to Brownian motion.

Doan et al. [2014] propose a multi-output Gaussian process model for misaligned time series, which is in contrast to other joint modeling work which assumes that the labels are aligned.[1] This is salient for Paleoclimate reconstruction as data will typically not be aligned, meaning that data products in which data has been "re-gridded" using some procedure are typically provided. This is potentially problematic if the data products contain only point estimates of the time series, as the point-estimate provided by the re-gridding procedure may smooth out important high-frequency information. The authors propose to alleviate this problem by providing a data product in which the point-estimates are replaced with posterior samples from their model. Note that they do not tackle the problem of uncertainty in the mapping from depth to time;

[1] Note that I'm still trying to figure out exactly what kernel they are using. They allude to the fact that it is linear, but the kernel is specified in terms of the variance of the difference between the random variables at two points on the process.

the authors acknowledge that it is a problem in general, but avoid the problem by considering data where there is good reason to believe that there is little uncertainty in this mapping.

Also of note is Tingley et al. [2012], which is an interesting position paper, and Haslett et al. [2006] provides the earliest example of the use of Bayesian inference for paleoclimate reconstruction that I have encountered.

To read: Tingley and Huybers [2010a], Tingley and Huybers [2010b], Li et al. [2007], Parnell et al. [2015] and Nieto-Barajas [2018], Smerdon [2012].

Definitely need to read: Lindgren et al. [2011], Arno's work (again).

3 Data Availability

[GHCN] is a general database for instrumental climate data. [NOAA] contains paleoclimate proxy data from various sources across the globe. [AGDC] provides cryospheric data primarily from the Antarctic, but there also appears to be date from Greenland. The data is freely available.

The European Pollen database Fyfe et al. [2009] appears to be more or less what it sounds like.

4 Some (Very) Preliminary Work

Figure 1 shows the results of modeling the deuterium data from Thomas et al. [2013] with a simple Gaussian process – it is assumed that we have access to noisy observations (red crosses) of a smooth latent process with an Exponentiated Quadratic kernel, where the noise is independent for each observation and has distribution $\mathcal{N}\big(0, \sigma^2\big)$. σ^2, and the length scale and variance of the EQ kernel were fitted using maximum marginal likelihood. The green crosses were not used during training to simulate missing data.

Qualitatively, it looks like we have found some structure, but it is hard to say much without further work.

References

Jason E Smerdon and Alexey Kaplan. Comments on "testing the fidelity of methods used in proxy-based reconstructions of past climate": The role of the standardization interval. *Journal of Climate*, 20(22): 5666–5670, 2007.

BAS. Ice cores and climate change. `https://www.bas.ac.uk/data/our-data/publication/ice-cores-and-climate-change/`, 2014. Accessed: 11-05-2018.

Arindam Banerjee and Claire Monteleoni. Climate change: Challenges for machine learning. `https://www-users.cs.umn.edu/~baner029/talks/BanerjeeMonteleoniNIPSTutorial2014.pdf`, 2014. Accessed: 11-05-2018.

Bo Li, Douglas W Nychka, and Caspar M Ammann. The value of multiproxy reconstruction of past climate. *Journal of the American Statistical Association*, 105(491):883–895, 2010.

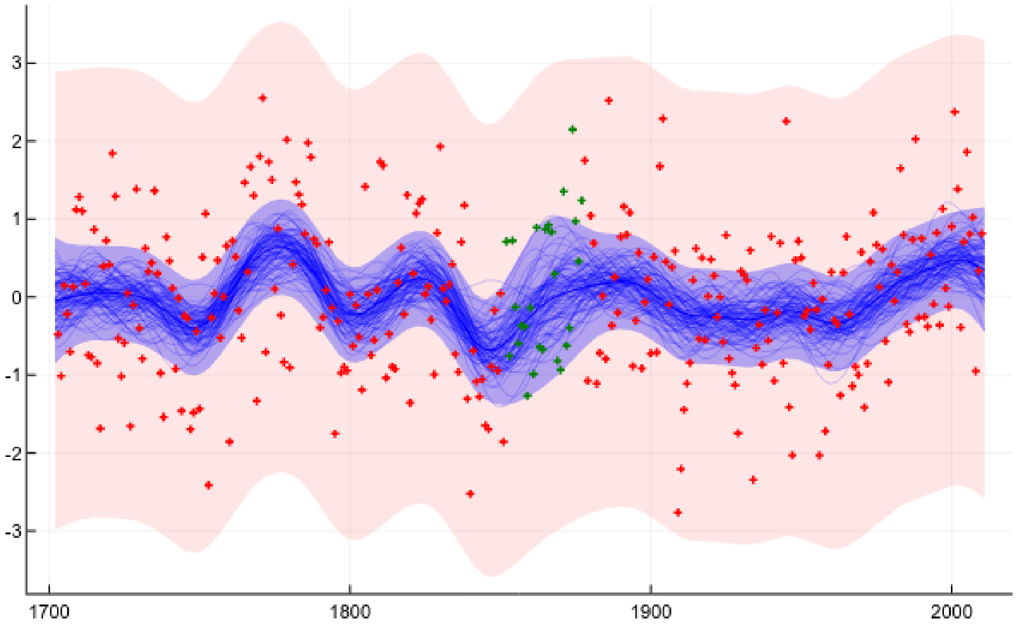

Figure 2: A simple Gaussian process fit to the Ferrigno deuterium data Thomas et al. [2013].

Blakeley B McShane and Abraham J Wyner. A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1000 years reliable? *The Annals of Applied Statistics*, pages 5–44, 2011.

Martin P Tingley and Peter Huybers. Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. *Nature*, 496(7444):201, 2013.

Thinh K Doan. *Bayesian Inference for Misaligned Irregular Time Series*. PhD thesis, Department of Statistics, Trinity College Dublin, 2015.

Thinh K Doan, Andrew C Parnell, and John Haslett. Joint inference of misaligned irregular time series with application to greenland ice core data. *arXiv preprint arXiv:1402.3014*, 2014.

Martin P Tingley, Peter F Craigmile, Murali Haran, Bo Li, Elizabeth Mannshardt, and Bala Rajaratnam. Piecing together the past: statistical insights into paleoclimatic reconstructions. *Quaternary Science Reviews*, 35:1–22, 2012.

John Haslett, Matt Whiley, Sudipto Bhattacharya, M Salter-Townshend, Simon P Wilson, JRM Allen, B Huntley, and FJG Mitchell. Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):395–438, 2006.

Martin P Tingley and Peter Huybers. A bayesian algorithm for reconstructing climate anomalies in space and time. part i: Development and applications to paleoclimate reconstruction problems. *Journal of Climate*, 23(10):2759–2781, 2010a.

Martin P Tingley and Peter Huybers. A bayesian algorithm for reconstructing climate anomalies in space and time. part ii: Comparison with the regularized expectation–maximization algorithm. *Journal of Climate*, 23(10):2782–2800, 2010b.

Bo Li, Douglas W Nychka, and Caspar M Ammann. The 'hockey stick'and the 1990s: a statistical perspective on reconstructing hemispheric temperatures. *Tellus A*, 59(5):591–598, 2007.

Andrew C Parnell, James Sweeney, Thinh K Doan, Michael Salter-Townshend, Judy RM Allen, Brian Huntley, and John Haslett. Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):115–138, 2015.

Luis Enrique Nieto-Barajas. Interpolation of paleoclimatology datasets. *Atmósfera*, 31(2):125–141, 2018.

Jason E Smerdon. Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. *Wiley Interdisciplinary Reviews: Climate Change*, 3(1):63–77, 2012.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

GHCN. Global historical climate network. `https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn`. Accessed: 11-05-2018.

NOAA. Paleo data search. `https://www.ncdc.noaa.gov/paleo-search/`. Accessed: 11-05-2018.

AGDC. Antarctic glaciolocal data center. `http://nsidc.org/data/agdc`. Accessed: 11-05-2018.

Ralph M Fyfe, Jacques-Louis de Beaulieu, Heather Binney, Richard HW Bradshaw, Simon Brewer, Anne Le Flao, Walter Finsinger, Marie-José Gaillard, Thomas Giesecke, Graciela Gil-Romera, et al. The european pollen database: past efforts and current activities. *Vegetation History and Archaeobotany*, 18(5): 417–424, 2009.

Elizabeth R Thomas, Thomas J Bracegirdle, John Turner, and Eric W Wolff. A 308 year record of climate variability in west antarctica. *Geophysical Research Letters*, 40(20):5492–5496, 2013.