Bayesian Palaeoclimate Inference from Pollen in Southern Italy

Andrew C. Parnell[1,2], James Sweeney[3], Thinh K. Doan[4],
Michael Salter-Townshend[2], Judy R.M. Allen[5], Brian Huntley[5],
and John Haslett[4]

[1]School of Mathematical Sciences (Statistics), University College Dublin, Ireland

[2]Complex and Adaptive Systems Laboratory, University College Dublin, Ireland

[3]School of Mathematical Sciences, University College Cork, Ireland

[4]School of Computer Science and Statistics, Trinity College Dublin, Ireland

[5]School of Biological and Biomedical Sciences, Durham University, UK

Abstract

We outline a model and algorithm to perform inference on the palaeoclimate and palaeoclimate volatility from pollen proxy data. We use a novel multivariate non-linear non-Gaussian state space model consisting of an observation equation linking climate to proxy data and an evolution equation driving climate change over time. The observation equation linking climate to proxy data is defined by a pre-calibrated forward model, created via a multivariate latent Gaussian process fitted via integrated nested Laplace approximation (INLA). The data for this forward model are taken from a calibration set of modern climate-pollen relationships. The evolution equation representing climate change is driven by a Normal-Inverse Gaussian Lévy process, being able to capture large jumps in multivariate climate whilst remaining temporally consistent. The pre-calibrated nature of the forward model allows us to cut feedback between the observation and evolution equations and thus integrate out the state nuisance parameters whilst making minimal simplifying assumptions. A key part of this approach is the creation of mixtures of marginal data posteriors representing the information obtained about climate from each individual time point. Our approach allows for an extremely efficient MCMC algorithm, which we demonstrate with a pollen core from Lake Monticchio in south-central Italy.

Keywords: Palaeoclimate Reconstruction, Normal-Inverse Gaussian Process, Marginal Data Posteriors, State-Space Models

1 Introduction

Palaeoclimate reconstruction is a major focus of the Intergovernmental Panel on Climate Change (Jansen et al., 2007). Public interest, however, has largely been fuelled by the 'Hockey Stick' and 'climategate' controversies, e.g. Mann et al. (1998, 1999); McShane and Wyner (2011) focussing on hemispheric climate changes over the past millennium. Such changes are relatively small and can be inferred with reasonable precision from proxies (e.g. tree rings) that are resolved annually. In contrast, the older Younger Dryas period (12.8ka to 11.5ka BP) shows a rapid switching from warm to cold to warm. During this period ice core data from Greenland show abrupt warmings of up to 16°C within decades (Jansen et al., 2007, P435). This size and rate of change is not captured well by the General Circulation Models (GCMs) which are used to predict future climate, nor by the precise proxies used to examine the past millennium. Pollen proxy data offer the best hope of resolving such sizeable past climate changes in locations other than Greenland.

The model we propose differs from many recent studies performing inference on palaeoclimate because:

(a) it is non-linear and non-Gaussian in the relationship between climate and proxy,

(b) we infer only climatic variables to which the proxy is sensitive,

(c) we use real data to produce climate reconstructions rather than simulated 'psuedo-proxies', and

(d) we allow for stochastic volatility in climate.

This approach, tackling many more sources of uncertainty than previously, is inspired by the SUPRANET group[1], of which many ideas are shared with Tingley et al. (2012) and Huntley (2012). The development of our modelling approach is only possible because a number of these aspects of palaeoclimate reconstruction have been fully developed by statisticians working in conjunction with palaeoclimate experts. We expand more on the importance of these various aspects in subsequent sections.

Our state-space model can be written as:

$$y_i|c_i \sim f_\theta(c_i, \mathcal{D}),\ i = 1\ldots, n \tag{1}$$

$$c_i = c_{i-1} + \gamma_i,\ i = 2, \ldots, n \tag{2}$$

where $y_i = y(t_i)$ represents pollen data at time t_i, $c_i = c(t_i)$ is palaeoclimate, $\mathcal{D}$ is a modern data set of pollen and climate ($\mathcal{D} = (c^m, y^m)$), f is a *forward model* parameterised by θ, and $\gamma_i \sim N(0, v(t_{i-1}, t_i))$ are innovations. Following convention, we call Equation 1 the observation equation, and Equation 2 the evolution equation. All of the parameters involved are multivariate: y_i being here of dimension 28 and c_i being of dimension 3. Our focus is on the marginal posterior distribution $\pi(c, v|y)$. One particular interest is in the distribution of the incremental variance terms $v_i = v(t_{i-1}, t_i)$, representing the square of the volatility. We set these to be Inverse Gaussian (Barndorff-Nielsen, 1997), written $IG(\eta, \phi)$, yielding a Normal-Inverse Gaussian (NIG) process on c.

The outline of this paper is as follows. In Section 2 we examine previous work in palaeoclimate reconstruction. In Section 3 we outline our modelling approach and a novel modular MCMC algorithm. We outline a case study focussing on climate change over the past 120k years in southern Italy in Section 4. We conclude in Section 5.

2 Previous work in statistical palaeoclimatology

There are three main model-based approaches which perform inference on palaeoclimate, all of which fit into our state-space model:

- Haslett et al. (2006) use 14 pollen taxa to reconstruct climate in two dimensions. In their approach, the forward model f is obtained from the climatological location of modern data in a multivariate spatial model where the proxy is the response. The evolution equation is represented by a t_8 random walk.
- Tingley and Huybers (2010) perform inference on mean annual temperature from psuedo-proxies (simulated data required to behave like proxy data). They use a linear observation equation, trained on the temporal overlap where both proxy and climate data are available. The evolution equation is represented by a fully spatial model with Exponential covariance and an MA(2) temporal component.
- Li et al. (2010) reconstruct mean annual temperature from multiple psuedo-proxies representing pollen, tree rings and borehole measurements. This model is similarly trained on temporally overlapping data, with a similarly linear observation equation. Here, however, the evolution equation is represented by an AR(2) model, and also allows for the inclusion of further covariates.

When the likelihood and prior distributions are Gaussian, as in Tingley and Huybers (2010) and Li et al. (2010) above, then Equations 1 and 2 are conceptually easy to solve via traditional MCMC or particle-type methods. Complications arise from the high-dimensionality of the climate parameters and the sheer quantity of data. Haslett et al. (2006), however, have a non-linear model and thus combine the MCMC methods with some approximations which partition the model into ‘modules’ (Liu et al., 2009). In particular, parameters θ of the likelihood are learnt solely from the modern data. Thus the fossil pollen information from any particular core contributes no further information. The modularisation

[1] Studying uncertainty in palaeoclimate reconstructions: http://caitlin-buck.staff.shef.ac.uk/SUPRAnet/

assumption is an unremarkable and implicit assumption in most studies involving instrumental data; e.g. experimental data on temperature are not normally used to re-calibrate the thermometer that generated them. In this paper we use a likelihood which is considerably more complex than that of Haslett et al. (2006), and thus similarly benefits from modularity assumptions.

A final common issue in the modelling of palaeoclimate data is the choice of climate variables to reconstruct. This choice, as outlined by Huntley (2012) should depend on the sensitivity of the proxy data to changes in climate. In contrast, many previous studies have chosen climate variables (e.g. mean annual temperature) which are not strongly affected by proxy changes. Huntley (2012) suggests:

- the Mean Temperature of the Coldest Month (MTCO) in Celsius, a measure of the harshness of the winter,
- the Growing Degree Days above 5°C (GDD5, also known as the annual temperature sum above 5°C), a measure of the warmth of the growing season,
- the ratio of Actual to Potential Evapotranspiration (AET/PET), a measure of the available moisture.

We include all three of these climate variables in our final model.

3 Statistical model

The three papers outlined at the start of Section 2 form the basis of our approach. We most closely follow that of Haslett et al. (2006) as this involves using real data and provides a fully non-linear model. We expand the number of pollen taxa from 14 to 28 and use the same spatial model to calibrate the observation equation. f is now a nested zero-inflated Binomial distribution, with θ including parameters which allow for zero inflation and measurement error. The modern data $\mathcal{D} = (y^m, c^m)$ represent 7742 modern samples of 28-dimensional pollen and 3-dimensional climate. Model fitting can be made much faster by treating the spatial model as a GMRF and so utilising the INLA framework of Rue et al. (2009). The forward modelling stage of our approach is more fully documented in Salter-Townshend and Haslett (2012).

We similarly extend the Haslett et al. (2006) approach to the evolution equation. We treat γ_i as Inverse Gaussian (IG) so that the climate process is marginally a Normal-Inverse Gaussian (NIG) distribution with volatilities $\sqrt{v_i}$. Bayesian inference for the NIG distribution has been discussed by Karlis and Lillestol (2004), providing closed-form complete conditionals and thus a neat Gibbs algorithm. Some of these complete conditionals transfer over to our model but others, when combined with the observation equation, are no longer required. The NIG process as we used it provide long-tailed climate behaviour, designed to match that experienced in the data.

The posterior distribution for our model can be written out as:

$$\begin{aligned}\pi(c, v, \eta, \phi, \theta | y, t, \mathcal{D}) \quad &\propto \quad \prod_{i=2}^{n} \pi(c_i | c_{i-1}, v_i) \prod_{i=1}^{n-1} \pi(v_i | t, \eta, \phi) \prod_{i=1}^{n} \pi(y_i | c_i, \theta) \\ &\quad \times \prod_{i=1}^{n^m} \pi(y_i^m | c^m, \theta)\ \pi(\theta)\ \pi(\eta, \phi) \end{aligned} \tag{3}$$

Such a model can be fitted using Markov chain Monte Carlo (MCMC) or particle methods (e.g. Carvalho et al., 2010; Andrieu et al., 2010). However, these tend to be extremely slow due to the high-dimensionality of the parameters c and v, as well as the large likelihood calculation of $\pi(y_i^m | c^m, \theta)$. We thus propose a modular structure to model fitting, following that proposed by Liu et al. (2009). Here, the proxy data y are not presumed to contribute any extra information to the posterior of θ. We can thus split the modelling into two modules; the first where we learn θ from the modern data (c^m, y^m), and the second where we learn c, v from the proxy data y.

In fact there are further benefits to be had following this approach. If parameters θ are learnt from a separate module, then it is trivial to create marginal data posteriors (MDPs) of the form $\pi(c_i|y_i)$ where we can marginalise over θ. By treating the MDP as a mixture of normal distributions, we end up with a posterior distribution which factorises in c, and so allows us to remove climate and focus inference directly on v. We report elsewhere on the detailed aspects of this fitting stage.

4 Case study: Lake Monticchio

Our case study covers a unique pollen core obtained from Lago Grande di Monticchio, discussed in Allen and Huntley (2009). The core contains 924 samples of pollen covering 132 thousand years before present. The creation of marginal data posteriors is undertaken on a layer by layer basis, and can be run in parallel. This step takes around 30 minutes on a Core-i7 processor with 16Gb Ram. The following MCMC run is much faster, taking just 2 minutes to perform 200,000 iterations, followed by the creation of interpolated climate histories and volatilities. We show marginal summaries of the climate histories for the three climate dimensions in Figure 1. We include the original MDPs and some output from the simpler method of Huntley (1993) for comparison.

The climate posterior distributions clearly show strong climate changes throughout the core's period with the strongest changes being seen in available moisture (AET/PET). Both the temperature based measurements (GDD5 and MTCO) seem relatively stable with only isolated periods of rapid change. The largest of these rapid changes occurs around 15k years BP; corresponding approximately with (though slightly earlier than) that of the Younger Dryas.

5 Discussion

Our approach allows for inference on palaeoclimate pollen cores with a more complete quantification of uncertainty than previously possible. In particular, the use of the NIG process as a prior distribution on climate change seems appropriate given the dynamic nature of the system. The fitting algorithm we have developed allows for fast inference on a high-dimensional complex model. The output from the model seems reasonable and broadly matches that given by other methods, though with a far greater focus on uncertainty. The algorithm is implemented in the R package `Bclim` and so is available for use by non-experts.

Future expansions of this model may allow for multiple cores in a spatial region to be run simultaneously. A spatial multivariate Normal-Inverse Gaussian process could be used, with an appropriate non-stationary spatial covariance structure. However, these will most likely have to be run on shorter timescales, as there are few (if any) which can match Monticchio for temporal coverage. Another extension would be to allow for multiple proxies to contribute simultaneously via independent forward models. However, these would need to be carefully matched for their temporal response to climate change can strongly vary. Such detailed forward models have yet to be built.

References

Allen, J. and B. Huntley (2009). Last Interglacial Palaeovegetation, Palaeoenvironments and Chronology: a New Record from Lago Grande di Monticchio, Southern Italy. *Quaternary Science Reviews 28*(15-16), 1521–1538.

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society - Series B: Statistical Methodology 72*(3), 269–342.

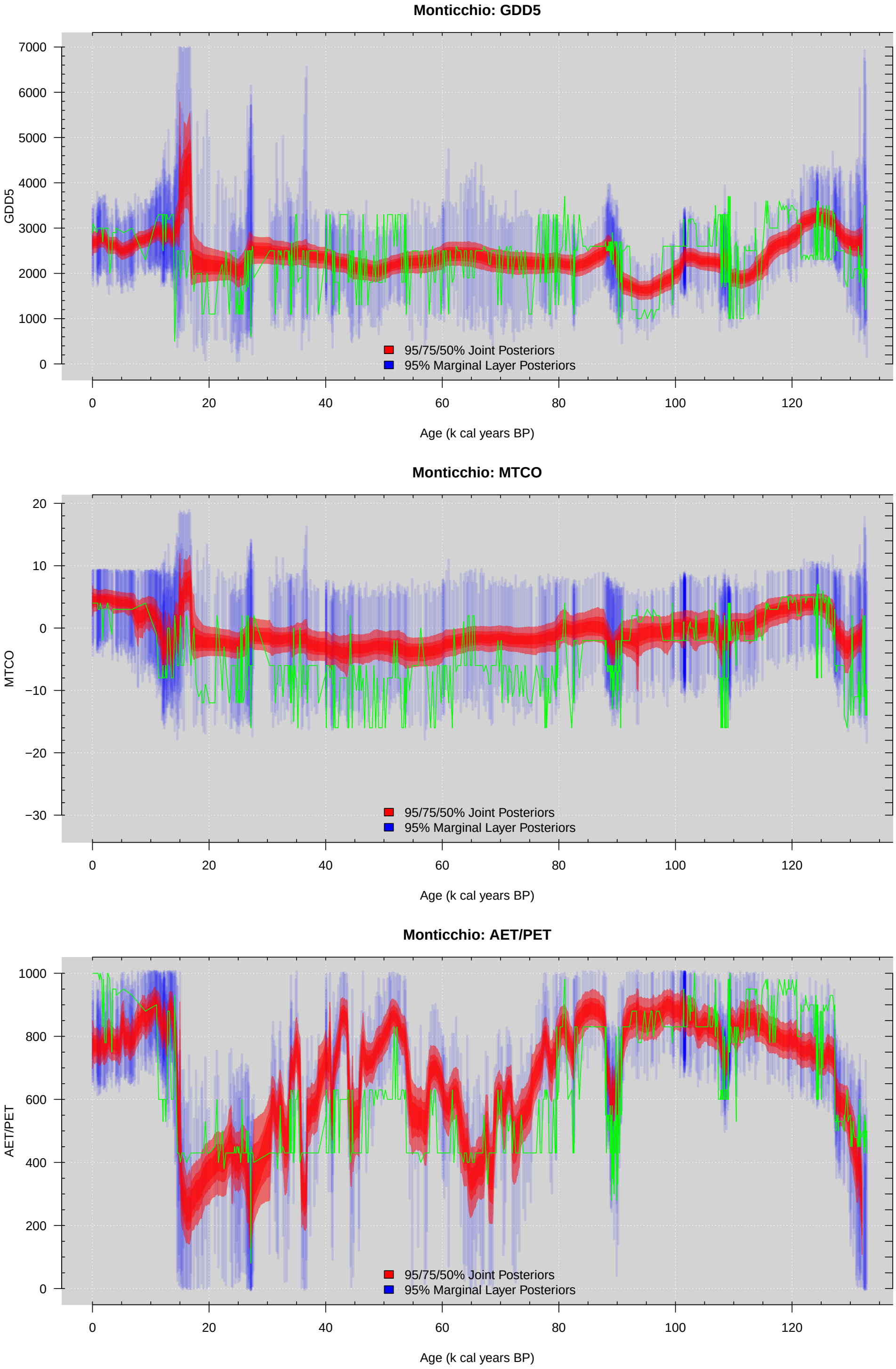

Figure 1: A plot of the centennial interpolated GDD5 (growing season warmth), MTCO (harshness of winter) and AET/PET (available moisture; scaled up to (0,1000)) over the period 0 to 125ka BP at Lake Monticchio. The blue 'blobs' represent the marginal data posteriors whereas the red bands represent summarised posterior stochastic interpolations of climates c. The green lines represent the output from the method of Huntley (1993).

Barndorff-Nielsen, O. E. (1997). Normal Inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics 24*(1), 1–13.

Carvalho, C. M., M. S. Johannes, H. F. Lopes, and N. G. Polson (2010). Particle learning and smoothing. *Statistical Science 25*(1), 88–106.

Haslett, J., M. Whiley, S. Bhattacharya, F. J. G. Mitchell, J. R. M. Allen, B. Huntley, S. P. Wilson, and M. Salter-Townshend (2006). Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society, Series A 169*, 395–438.

Huntley, B. (1993). The use of climate response surfaces to reconstruct palaeoclimate from Quaternary pollen and plant macrofossil data. *Philosophical Transactions of the Royal Society of London, Series B - Biological sciences. 341*, 215–223.

Huntley, B. (2012). Reconstructing palaeoclimates from biological proxies: some often overlooked sources of uncertainty. *Quaternary Science Reviews 31*, 1–16.

Jansen, E., J. Overpeck, K. R. Briffa, J.-C. Duplessy, F. Joos, V. Masson-Delmotte, D. Olago, B. Otto-Bliesner, W. R. Peltier, S. Rahmstorf, R. Ramesh, D. Raynaud, O. Rind, R. Solomina, V. R., and D. Zhang (2007). Palaeoclimate. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, T. M., and H. L. Miller (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Karlis, D. and J. Lillestol (2004). Bayesian estimation of NIG models via Markov chain Monte Carlo methods. *Applied Stochastic Models in Business and Industry 20*(4), 323–338.

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

Liu, F., M. J. Bayarri, and J. O. Berger (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis 4*(1), 119–150.

Mann, M. E., R. S. Bradley, and M. K. Hughes (1998). Global-scale temperature patterns and climate forcing over the past six centuries. *Nature 392*(6678), 779–787.

Mann, M. E., R. S. Bradley, and M. K. Hughes (1999). Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophysical Research Letters 26*(6), 759.

McShane, B. B. and A. J. Wyner (2011). Discussion of: A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable? *The Annals of Applied Statistics 5*(1), 1–45.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B 71*, 1–35.

Salter-Townshend, M. and J. Haslett (2012). Fast inversion of a flexible regression model for multivariate pollen counts data. *Environmetrics 23*(7), 595–605.

Tingley, M. P., P. F. Craigmile, M. Haran, B. Li, E. Mannshardt, and B. Rajaratnam (2012). Piecing together the past: statistical insights into paleoclimatic reconstructions. *Quaternary Science Reviews 35*, 1–22.

Tingley, M. P. and P. Huybers (2010). 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.