Random-walk radiocarbon calibration

J Andres Christen
Instituto de Matematicas
Campus Morelia, UNAM
Morelia, Michoacan, Mexico
jac@matmor.unam.mx

Geoff Nicholls
Department of Mathematics
Auckland University
Auckland, New Zealand
nicholls@math.auckland.ac.nz

June 27, 2000

Abstract

In this note we check the interesting recent results of Gomez Portugal Aguilar et al [1]. These authors fit a simple random walk with Gaussian increments to the radiocarbon calibration curve. They suggest that the posterior standard deviation of the calibration curve is smaller between years where observations were made than it is at those years. We find that in contrast to their result the posterior standard deviations of the calibration curve bow out slightly between observation points in our estimates, for a model of the kind they describe. In earlier work Christen [2] conditions the random walk to visit the measured calibration values. Following [1], we consider the unconditioned posterior distribution for calibration curves, and thereby improve slightly on the reconstructions of [2].

1 Introduction

We consider two data sets, the University of Washington 1998 single year 14C dataset [3] covering years 1511 up to 1954 and the 1998 decadal atmospheric delta 14C and radiocarbon age dataset [4] called INTCAL98. We wish to construct the posterior distribution for the radiocarbon calibration curve obtained by taking a random walk prior with Gaussian increments for the calibration function, combined with the data in these two sets. This choice of prior has rather weak physical motivation and such unphysical properties as non-smooth sample paths and unbounded variance. An integrated Ornstein-Uhlenbeck process makes a more credible prior. However, as an integrated Ornstein-Uhlenbeck process is rather cumbersome to analyse, we put it aside for the moment.

2 Prior and observation models

For $L, U \in \mathbb{Z}$, let $[L, U]$ denote the discrete set of years from L to U inclusive. Let $M(t)$ record the unknown true value of the radiocarbon calibration curve at calendar year $t \in [L, U]$. Our prior model for M is a random walk with Gaussian increments,

$$M(t+1) \sim N(M(t) + E, \Sigma^2),$$

approximating simple 1-D Brownian motion over large time scales. Here Σ^2 is the unknown true variance of increments in the random walk process and E their mean. We take "free" boundary conditions for $M(L)$ and $M(U)$. Our prior is improper as curves M and $M + c$ for c constant have the same prior density.

We condition on zero drift, setting the drift parameter $E = 0$, as all aspects of reconstructions are insensitive to this parameter. This may seem surprising until we realise that the conditional distribution of $M(t)|M(t-1), M(t+1)$ is independent of the drift parameter. Drift terms are important at boundaries, where the densities of $M(L)|M(L+1)$ and $M(U)|M(U-1)$ are functions of the drift parameter. However, in the presence of data of the kind we have here, events at the boundary do no propagate into the centre of the reconstructed interval in any significant way. Simulations (not reported) show that conditioning on zero drift inflates the posterior mean for Σ by not more than a few percent for the single year data.

Let D denote the set of dates for which observations of M are available. Let $y = \{y(t), t \in D\}$ label this set of observations. Let $s = \{s(t), t \in D\}$ be a set of corresponding standard errors. The observation model for a realisation $y(t)$ of the observation process $Y(t)$ is Gaussian

$$Y(t) \sim N(M(t), s(t)^2),$$

iid for each $t \in D$, with $s(t)$ assumed known exactly. This model is incorrect insofar as observations are in fact averages over blocks of years about the given year. We ignore this in a first pass, though it is straightforward to improve the observation model in this respect, when inference is sample-based.

Let μ be some trial curve for M, *ie* μ is a vector with entries $\mu = (\mu(L), \mu(L+1) \dots \mu(U))$. Let σ denote a trial value for Σ. We take the improper prior density $1/\sigma^2$ for Σ. The posterior density for μ and σ has the form

$$h_{M,\Sigma}(\mu, \sigma | y, s) \propto f_Y(y|\mu, s) f_M(\mu|\sigma) f_\Sigma(\sigma),$$

where

$$f_Y(y|\mu, s) \propto \prod_{t \in D} e^{-(y(t)-\mu(t))^2/2s(t)^2} / s(t),$$

is our likelihood,

$$f_M(\mu|\sigma) \propto \sigma^{-U+L} \prod_{t \in [L, U-1]} e^{-(\mu(t+1)-\mu(t))^2/2\sigma^2}$$

is the prior density for the calibration curve determined by our choice of a random walk prior, and

$$f_\Sigma(\sigma) \propto 1/\sigma^2,$$

is our prior density for Σ, the one remaining hyper-parameter.

3 Proposed inference

It is sufficient, for our purposes, to proceed in the following way. Using the single year data set [3], we estimate the posterior distribution of Σ using MCMC simulation. We simulate a sequence $\{\mu^{(n)}, \sigma^{(n)}\}_{n=1}^{J}$ of J samples from $h_{M,\Sigma}(\mu, \sigma | y, s)$ and simply ignore the $\mu^{(n)}$ component.

We are assuming Σ is constant in time. We test this assumption by repeating the analysis on the single year data taken in one hundred year blocks. We see a small amount of real variation between blocks but no consistent trend in the posterior distribution of Σ (which would be more of a worry). From these simulations we obtain a point estimate $\hat{\sigma} = 7.9$ for Σ. This one number represents the marginal posterior distribution for Σ shown in Figure 2.

We then take the INTCAL98 data of [3] and simulate the posterior for the calibration curve M, conditioning on $\Sigma = 7.9$. We expect that the INTCAL98 data contains very little information about Σ in comparison to the single year data set. Also, we are interested in the qualitative behaviour of the reconstructed calibration curve, and we expect this to be insensitive to the details of the distribution of Σ, if we are conditioning on a reasonably central value of that parameter.

A simple MCMC simulation scheme was used, and coded in MATLAB and independently in C, with identical results. We report simulations from an interval of 200 years. Variances at the edge of the interval are inflated by the missing data outside the interval. However, since correlations along the curve are weak, the behaviour of our reconstructions in the centre of the interval are representative of the results which would be obtained from the full data set. It is practicable, in our C implementation, to reconstruct M over the high precision interval of INTCAL98 from 0 to 6000 BP. Such reconstructions will be available from the authors.

4 Simulation details

Let $\mu^{(n)}, \sigma^{(n)}$ record the state at the n-th update in the MCMC simulation. The μ-update is Gibbs, whilst the σ-update is random-walk Metropolis Hastings. At an update of μ, a time t is selected uniformly at random from $[L, U]$, and a new value for $\mu(t)$ sampled conditional on all other details of the state. For the case where t is not equal L or U, and $t \notin D$, the new value is simply

$$\mu^{(n+1)}(t) \sim N((\mu^{(n)}(t-1) + \mu^{(n)}(t+1))/2, {\sigma^{(n)}}^2/2).$$

If we select a year for which there is data, so $t \in D$, the conditional mean is

$$m^{(n)} = \frac{s(t)^2 \left(\mu^{(n)}(t-1) + \mu^{(n)}(t+1)\right) + {\sigma^{(n)}}^2 y(t)}{2s(t)^2 + {\sigma^{(n)}}^2}$$

and the conditional variance is

$${v^{(n)}}^2 = \frac{1}{2/{\sigma^{(n)}}^2 + 1/s(t)^2}$$

so that we sample

$$\mu^{(n+1)}(t) \sim N(m^{(n)}, {v^{(n)}}^2)$$

to get a Gibbs sampler. The formulae for the special cases $t = L$ and $t = U$ are likewise straightforward to write down.

At an update of σ, a candidate value σ' is chosen uniformly at random in a window of fixed width $2d_\sigma$ centred on $\sigma^{(n)}$ ($d_\sigma = 2$ gave a reasonable mixing rate for the MCMC). The candidate is accepted (so $\sigma^{(n+1)} = \sigma'$) with probability $\min\{1, r_\sigma\}$ where

$$\log(r_\sigma) \quad = \quad (U - L + 2)\log(\sigma^{(n)}/\sigma') \quad + \quad \frac{1}{2}(1/{\sigma^{(n)}}^2 - 1/{\sigma'}^2) \sum_{t \in [L, U-1]} (\mu(t+1) - \mu(t))^2.$$

If the candidate state is rejected we set $\sigma^{(n+1)} = \sigma^{(n)}$.

In our MCMC simulation of $M, \Sigma | y, s$ for the single year data (Figures 1 and 2), we simulated 23144 updates, taking a sample at each update. The autocorrelation time τ_Σ proved to be $\tau_\Sigma = 39(4)$ updates, so our correlated MCMC output was equivalent to about 600 independent realisations of $M, \Sigma | y, s$. Note that the autocorrelation we are talking about here, which runs along the MCMC update label n and determines the accuracy of statistics estimated from MCMC output, is not to be confused with the autocorrelation of the next section which runs along years t, and is one of our summary statistics for the posterior distribution for $M | y, s, \Sigma$.

In our MCMC simulation of $M | y, s, \Sigma$ for the INTCAL98 data (Figures 3, 4 and 5) we simulated 59448 updates, taking a sample each 10 updates. We report the autocorrelations in the MCMC output of variable $M(1250)$. The autocorrelation time $\tau_{M(1250)}$ proved to be $\tau_{M(1250)} = 109(16)$ updates, so our MCMC output was equivalent to about 550 independent realisations of $M | y, s, \Sigma$.

5 Results

The results of our simulations are shown in the following figures. All error bars are 1σ. Two realisations from the posterior distribution of $M | y, s, \Sigma$ for the INTCAL98 data set are plotted in Figure 3. Let $\hat{\mu}(t)$ be the standard estimator for $\mathsf{E}_M\{M(t) | y, s\}$ formed from MCMC output and let $\hat{\sigma}_M(t)^2$ be the corresponding estimator for $\mathrm{var}(M(t)) = \mathsf{E}_M\{(M(t) - \mathsf{E}\{M(t)\})^2 | y, s\}$. Estimates of $\hat{\mu}(t)$ and $\hat{\sigma}_M(t)$ were made for each $t \in [L, U]$. We plot $\hat{\mu}(t)$, and $\hat{\mu}(t) \pm \hat{\sigma}_M(t)$ estimated from the single year data in Figure 1. The same kind of information is represented in Figure 4, however this time we are interpolating a section of the INTCAL98 data and conditioning on $\Sigma = 7.9$.

Notice that, as Gomez Portugal Aguilar *et al* point out in [1], the posterior variance of the calibration is reduced at measurement years, so that, for $t \in D$,

$$\mathsf{E}_M\{(M(t) - \mathsf{E}\{M(t)\})^2 | y, s, \Sigma\} \leq s(t)^2.$$

However, in contrast to Gomez Portugal Aguilar *et al*, we find, as we would expect, that the posterior variance of $M(t)$ increases as the time interval to the nearest measurement year increases.

When we summarise the posterior distribution for M using estimates, $\hat{\mu}(t)$ and $\hat{\sigma}_M(t)$, of its year by year means and variances, we hide the correlations which exist between $M(t)$ and $M(t+\tau)$, for τ an integer offset. In Figure 5 we display the estimated autocorrelation function

$$\rho(\tau, t) = \frac{\mathrm{cov}(M(t), M(t+\tau))}{\mathrm{var}(M(t))}$$

for $M(t)$, choosing a year $t = 1250$ which lies in the centre of the reconstructed interval. The correlation between $M(t)$ and $M(t+\tau)$ falls off to zero over about twenty years, supporting our earlier assertion that boundary effects are unimportant here. The correlations we see here may be worth taking into account, through a revised expression for the likelihood, when a number of radiocarbon dates fall in a small interval. Such datasets are common in New Zealand, as the prehistory is comparatively short.

References

[1] D. Gomez Portugal Aguilar, C. D. Litton, and A. O'Hagan. A new piece-wise linear radiocarbon curve wih more realistic variance. Technical report, http://www.shef.ac.uk/ st1ao/ps/rep495.ps, Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, UK, 2000.

[2] J. A. Christen. *Bayesian Interpretation of Radiocarbon Results*. Phd thesis, University of Nottingham, 1994.

[3] M. Stuiver, P. J. Reimer, and T. F. Braziunas. High-precision radiocarbon age calibration for terrestrial and marine samples. *Radiocarbon*, 40:1127–1151, 1998.

[4] M. Stuiver, P. J. Reimer, E. Bard, J. W. Beck, G. S. Burr, K. A. Hughen, B. Kromer, F. G. McCormac, J. v. d. Plicht, and M. Spurk. INTCAL98 radiocarbon age calibration, 24,000-0 cal BP. *Radiocarbon*, 40:1041–1083, 1998.

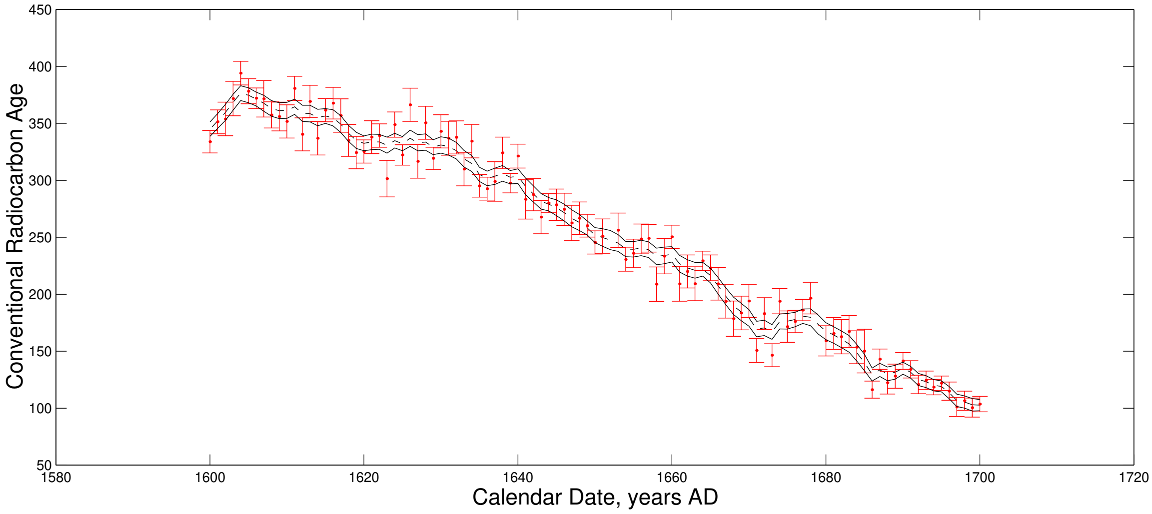

Figure 1: Single year data. Part of the posterior mean calibration curve $\hat{\mu}(t)$ with posterior standard deviation $\hat{\sigma}_M(t)$ of the calibration curve itself (1σ). We emphasise that we are not displaying confidence intervals for the estimated mean. We are summarising the distribution of $M(t)$ not the distribution of $\hat{\mu}(t)$.

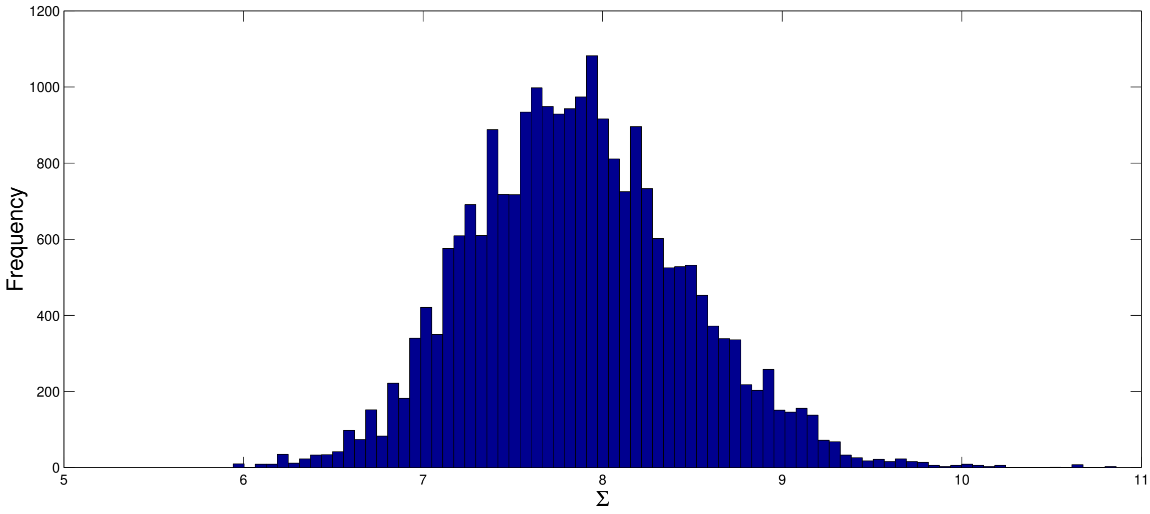

Figure 2: Marginal posterior distribution of the instantaneous variance Σ of the random walk increments for the single year data

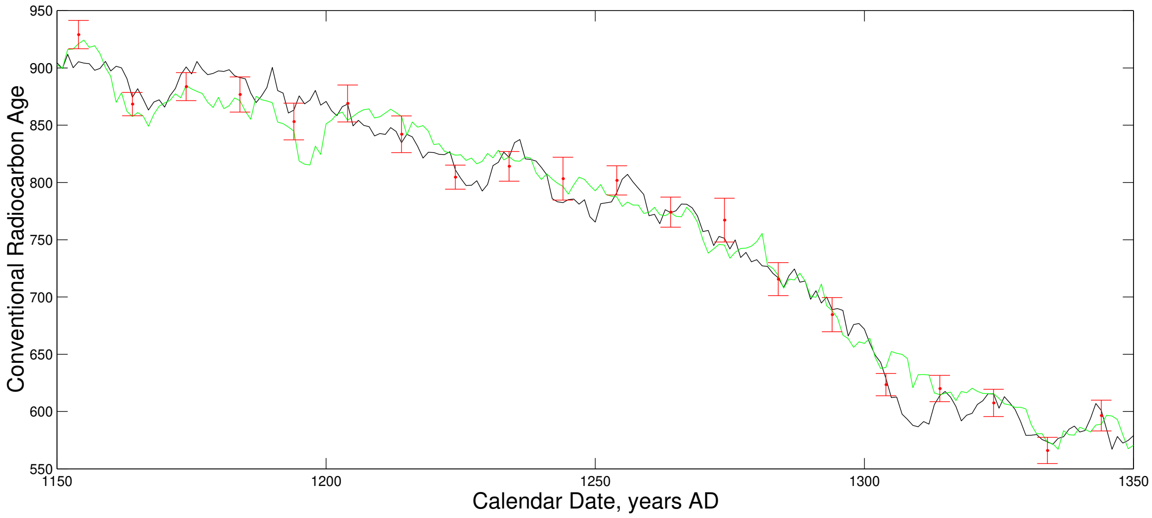

Figure 3: Two realisations of $M|y, s, \Sigma$ (solid lines). These are typical of the calibration curves interpolating the INTCAL98 calibration data (error bars, 1σ), for a section of the calendar scale, in a random walk model of the calibration function. This simulation conditions on variance $\Sigma^2 = 7.9^2$ for the random walk increments.

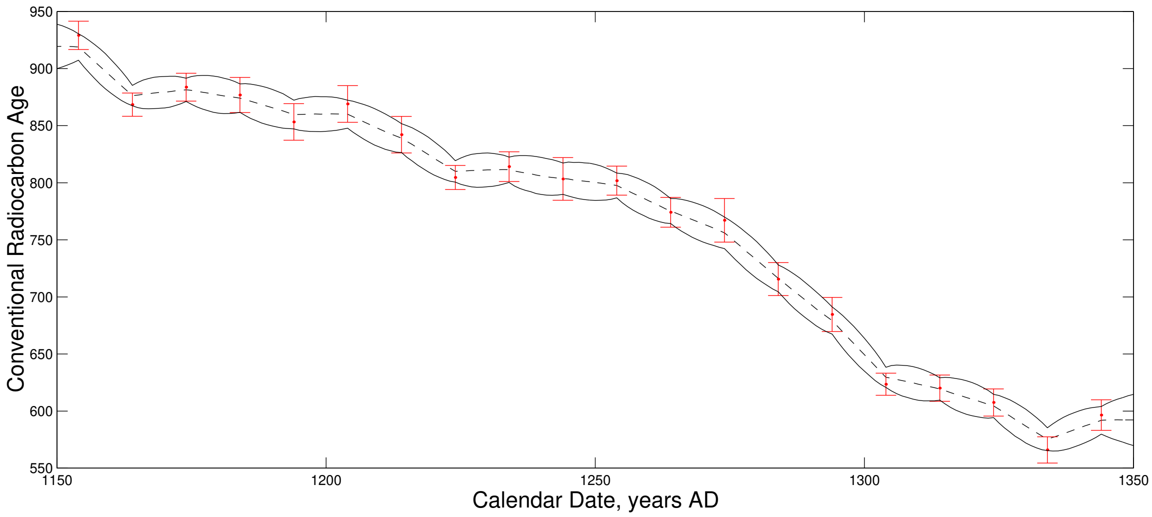

Figure 4: Posterior mean calibration curve $\hat{\mu}(t)$ (dashed) with posterior standard deviation $\hat{\sigma}_M(t)$ (solid lines, 1σ) of the calibration curve itself (1σ), estimated for each year from $L = 1150$ up to $U = 1350$. The data set (error bars, 1σ) is INTCAL98, and the simulation conditions on variance $\Sigma^2 = 7.9^2$ for the random walk increments.

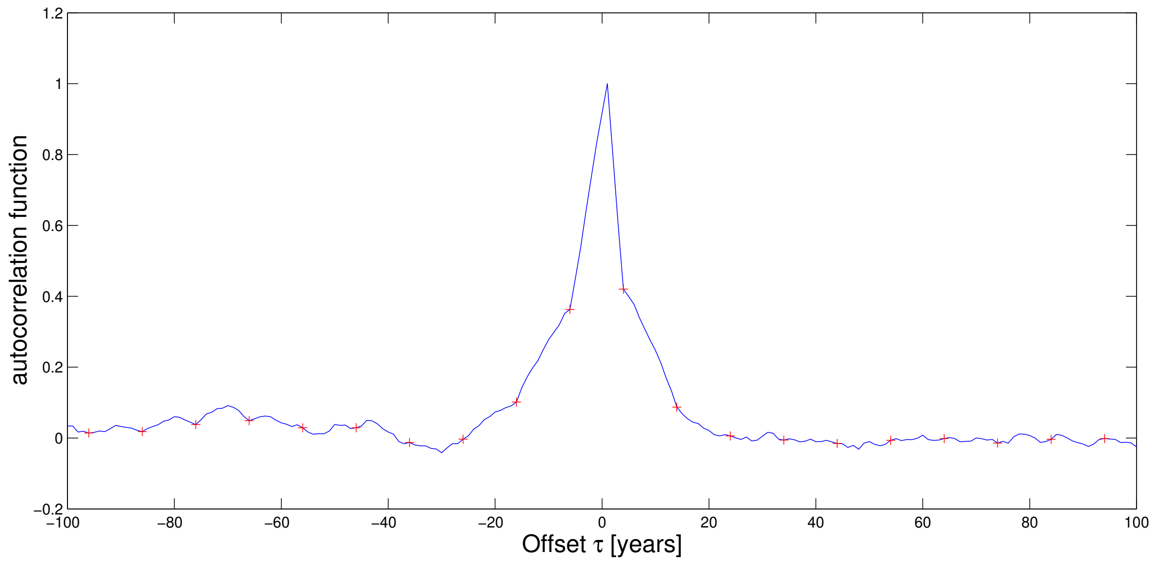

Figure 5: Posterior autocorrelation $\rho(\tau, t)$ for $M(t)$ with t the year 1250. The data set is INTCAL98, and the simulation conditions on variance $\Sigma^2 = 7.9^2$ for the random walk increments. The $+$ symbols mark τ for which $1250 + \tau \in D$, *ie* a datum occurs in Figure 4.