Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models
Kerry Gallagher, Thomas Bodin, Malcolm Sambridge, Dominik Weiss, Malin Kylander, David Large

To cite this version:
Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models

Kerry Gallagher¹, Thomas Bodin², Malcolm Sambridge², Dominik Weiss³,
Malin Kylander⁴, David Large⁵

1 – Géosciences, Université de Rennes 1, Rennes, 35042, France
2 – Research School of Earth Sciences, Australian National University, Canberra, ACT 0200, Australia
3 – Dept. of Earth Science and Engineering, Imperial College London, London SW7 2AZ, England
4 – Dept. of Geological Sciences, Stockholm University, 106 91 Stockholm, Sweden
5 – Dept of Chemical and Environmental Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, England

Corresponding author :
Kerry Gallagher, kerry.gallagher@univ-rennes1.fr, +33 (0)2 23 23 6081
Abstract

We present a method to quantify abrupt changes (or changepoints) in data series, represented as a function of depth or time. These changes are often the result of climatic or environmental variations and can be manifested in multiple datasets as different responses, but all datasets can have the same changepoint locations/timings. The method we present uses transdimensional Markov chain Monte Carlo to infer probability distributions on the number and locations (in depth or time) of changepoints, the mean values between changepoints and, if required, the noise variance associated with each dataset being considered. This latter point is important as we generally will have limited information on the noise, such as estimates only of measurement uncertainty, and in most cases it is not practical to make repeat sampling/measurement to assess other contributions to the variation in the data. We describe the main features of the approach (and describe the mathematical formulation in supplementary material), and demonstrate its validity using synthetic datasets, with known changepoint structure (number and locations of changepoints) and distribution of noise variance for each dataset. We show that when using multiple data, we expect to achieve better resolution of the changepoint structure than when we use each dataset individually. This is conditional on the validity of the assumption of common changepoints between different datasets. We then apply the method to two sets of real geochemical data, both from peat cores, taken from NE Australia and eastern Tibet. Under the assumption that changes occur at the same time for all datasets, we recover solutions consistent with those previously inferred qualitatively from independent data and interpretations. However, our approach provides a quantitative estimate of the relative probability of the inferred changepoints, allowing an objective assessment of the significance of each change.

Keywords: Transdimensional changepoint models, Geochemical data, Bayesian modelling, climate change
1. Introduction

A major issue in the interpretation of geochemical data (represented as depth or time series) is the detection of changes in trends. Reliable identification of short-term variations superimposed on long-term trends is critical for answering questions about uniformity in the rates of geological processes. The research effort towards understanding climate change often focuses on the inference of rapid or abrupt changes in the mean signal over time. Such environmental changes are recorded by geochemical proxies (e.g. McDermott 2004, Carslaw et al. 2006, Kylander et al. 2007, 2009, 2010, Gutjahr et al 2008, Cloy et al. 2008, Yasahuru et al. 2008, Ruggieri et al 2009, Large et al 2009, Cole et al. 2009, Palmer et al. 2010, Burton et al. 2007, 2010). Ideally, recognition of signals from data should based on sound qualitative interpretation, but also involve quantitative inference from the observations, allowing for possibly unknown noise levels in the data.

Here we recognise that the definition of abrupt or rapid is subjective. The definition and identification of change also depends on the form of the trend we expect between changes. We define abrupt changes as statistically significant variation in the trend over a scale of one or two samples of the total dataset. Furthermore, we note that the conversion of depth to time (or age) generally involves calibration of a depth-age relationship, which itself will have uncertainty (e.g. Thompson and Goldstein 2006, Kylander et al. 2009, 2010). For data collected from one borehole for example, these uncertainties will not change the positions of the underlying changes, only affecting the inference of absolute timing, but may become important for the inference of simultaneous changes in data from different locations.

To provide a brief overview of some approaches for inferring abrupt changes in geochemical records, we draw on a selection of published work, including some of our own. Large et al. (2009) described geochemical data including C, N, H concentrations and
C, H and O isotopes from a 6 m core taken in the Hongyuan peatland, eastern Qinghai–
Tibetan Plateau. The core material was dated back to 9.6 kiloyears using $^{14}$C. The
interpreted palaeoenvironmental history was linked to the climate variations in northwest
Pacific, the El Niño-Southern Oscillation, movement of the Intertropical Convergence Zone
and the East Asian Monsoon. This interpretation was based on a qualitative visual
inspection and comparison to other proxy data and interpretations.

Ruggieri et al. (2009) developed a method to infer Milankovitch–type cycles from
geochemical ($\delta^{18}$O) data. They allow for discrete changes (changepoints) between which
the trend (defined by superimposed sine functions) can change abruptly. They apply their
model to 2 sets of benthic $\delta^{18}$O isotope data with time ranges going back to 2500 and 5000
kiloyears. As these authors state, an important limitation of their method is that they do not
include the number of change points as a parameter to be inferred directly. Instead they
examine the variation of the data fit as a function of the number of changepoints, and try to
identify the upper limit such that adding more changepoints makes little difference to the
data fit. One additional limitation of this approach is that the inference will depend on the
errors inherent in the data, although the data fit function adopted by Ruggieri et al (2009)
does not incorporate a data error term explicitly. In general, we expect a more complex
model (i.e. more changepoints) given more precise data. Often however, we do not have
reliable estimates of the data errors and then choosing a suitable model becomes an issue.

Tomé and Miranda (2004), looking for changes in linear trends, fit gradients to time
series, subject to constraints on the minimum distance between changepoints, and the
magnitude of the changes in trends. The approach requires a user to specify a range for
number of changepoints, and finds best fitting functions for each value of the number of
change points in turn. Although the authors state that they can examine a series of the sums
of squares of residuals for each model (with different numbers of changepoints) to choose
an appropriate value, they do not explicitly do this. Rather they seem to favor visual
inspection to select a preferred model, which is required to have a constant distance
between the changepoints.

Finally, Kylander et al. (2007) analyzed rare earth elements and lead isotopes from
samples in an Australian peatland as inorganic proxies for climatic variations, reflected in
atmospheric dust. The changepoint modelling approach described in Denison et al. (2002)
was adapted to allow common changepoints for multiple datasets. Kylander et al. (2007)
used Eu abundance and Pb isotopes to infer the changepoints, under the assumption that
trends in the data between the changepoints could be expressed as a simple linear
regression function (a constant value or constant slope). In contrast to previous approaches,
the number of changepoints was a parameter to be inferred directly. Moreover, the problem
is formulated in a Bayesian framework, and provides probability distributions on the
number of changepoints and on the changepoint locations.

In this paper, we present a general approach to infer an unknown number of
changepoints when the errors, or the noise variances, are unknown for one or more datasets.
The data can be irregularly spaced in depth (or time) and there is no requirement for
different datasets to be sampled at the same depths (time). We specify nothing about the
spacing between changepoints, and the approach we present generalizes to any linear
function between changepoints. A key difference in the approach we present here to that of
Denison et al. (2002) is that we decouple the estimates of the data noise variance and the
model parameters. A related approach to deal with poorly constrained errors was described
by Malinverno and Briggs (2004) applied to seismic travelttime inversion for 1D velocity
structure (and see also Malinverno and Parker, 2006).

We begin with a general discussion of changepoint modelling and the role of
different sources of variability in the data. Then we describe the approach we adopt,
although most of the mathematical details are given in the supplementary material. The method we present includes not only estimation of the number and location of changepoints, the regression functions between changepoints, but also the distribution of the data noise variance (if it is unknown or considered unrepresentative). We conclude with some examples of the application of the method to synthetic and real data and a brief discussion/summary.

2. Changepoint modelling and noise

The general problem can be stated as follows (and see figure 1); given one or more sets of (geochemical) data, $f(x_i)$, $i = 1,N$, at positions $x_i$, representing a depth or time series, with either known or unknown levels of noise, can we identify the underlying trends or signal (e.g. the mean or a more general regression function) and the locations of discrete changes in the trends? Often, we may want to infer the same changepoint locations, but with different signals in each dataset.

In general we do not know how many changes are appropriate and ideally we should estimate this from the data. The problem becomes a transdimensional inverse problem (that is we do not specify in advance the number of unknown parameters, e.g. Sambridge et al. 2006). Furthermore, there is a trade-off between the level of noise in the data, and how well we expect to fit the data. Here we broadly follow the philosophy of Scales and Snieder (1999) in considering noise as that part of the data that we do not wish/expect the model to explain. The spread in a geochemical dataset can then be divided into the variation ($\sigma_{GP}^2$) due to time/depth varying geological processes (which we are interested in understanding) and the variation due to geological ($\sigma_{GN}^2$) and analytical ($\sigma_{AN}^2$) noise. Here geological noise might arise from spatial or temporal variations in small scale
or short term processes, local geological/biological/atmospheric variability and analytical
noise (or more commonly, errors) typically can arise from factors such as instrumental drift
or calibration from imperfect standards. In general we can express the total variance, \( \sigma_T^2 \) as

\[
\sigma_T^2 = \sigma_{GP}^2 + \sigma_{GN}^2 + \sigma_{AN}^2
\]  

(1)

When we are interested in identifying trends and changes in trends, we would like to
choose a model that adequately explains the variation due to geological process, and then
the residuals between the model predictions and the observed data reflectsthe 2 sources of
noise as defined above. We want good control on the noise (or noise variance) as this will
directly influence how well we should fit our data. Intuitively, we can see that if we
consider a series of scattered data with lower noise variance than is appropriate then we
will tend to fit many changepoints. An extreme case would be if we assume no noise, then
we will fit the data perfectly (with a changepoint between each data point). On the other
hand if we regard the data as more noisy than they really are, then we will tend to fit a
model that has too few changepoints.

2.1 Bayesian formulation of the change-point modelling problem

Underpinning the Bayesian approach is that unknowns are expressed in terms of probability
density functions (e.g. Tarantola and Valette 1982). A common form of Bayes’ rule is

\[
p(m|d) \propto p(d|m) p(m)
\]  

(2)

where \( p(m|d) \) is the probability density function (PDF) of the unknown model parameter
vector, \( m \), containing the unknowns, given the data vector, \( d \); \( p(d|m) \) is the likelihood
function which is effectively the probability of the data, \( d \), being observed given the model, \( m \). The likelihood increases as the model fits the data better relative to the data noise and the form of the likelihood depends on the statistical character of the noise on the data. Finally, \( p(m) \) is the prior PDF on the model (that is what we think we know about \( m \) before we have the data). The aim of Bayesian inference is to try and estimate the posterior PDF, \( p(m|d) \) as this characterises all we need to know about the distribution of model parameter values, given the prior and the information contained in the data (incorporated through the likelihood). Useful references for Bayesian inference are by Box and Tiao (1973), Lee (1989), Bernardo and Smith 1994) and Gelman et al (2004).

2.1.1 Model parameters

In a changepoint problem, the unknown model parameters are the number of changepoints \( (n) \), their locations \( (c) \), the parameters of a regression function between the changepoints \( (A) \) and the noise level \( (\sigma) \) for each dataset being considered. Thus we can write a general model vector, \( m \), as

\[
m = (n,c,A,\sigma)
\]

in which \( c, A, \) and \( \sigma \) can all be vectors.

We write the unknown locations of changepoints as \( c_i, i = 1,n \) (note that \( n \) itself is also an unknown). We refer to the region between each changepoint as a partition, and a predictive regression function, \( f_i(x) \), is defined whose parameters \( (A) \) depend on the data in that partition. Thus \( f_i(x) \) refers to the regression function within the partition at the left of change point \( c_i \) (if these are considered on a horizontal axis). As there are \( n+1 \) partitions for \( n \) change points, we define \( c_0 \) as the location of the first data point so the regression function \( f_1 \) in the first partition is defined between \( c_0 \) and \( c_1 \).
Given vectors of independent and dependent (observed) variables, \(x\) and \(d_{\text{obs}}\), such as depth (or time) and observed data respectively, the linear regression function between changepoints can be written as

\[
f(x) = \sum_{i=1}^{M} \alpha_i G_i(x)
\]

(3)

where \(G_i\) represents a specified basis function and \(\alpha_i\) represents an unknown coefficient, which can be thought of as weights on each basis function. For example, the common straight line relationship given by \(d_{\text{pred}} = \alpha_1 + \alpha_2 x\), is written as a vector-matrix equation,

\[
d_{\text{pred}} = G^t A
\]

where the superscript \(t\) represents the matrix transpose, and

\[
G = \begin{pmatrix}
1 & x_1 \\
\vdots & \vdots \\
1 & x_i \\
\end{pmatrix}, \quad A = \begin{pmatrix}
\alpha_1 \\
\alpha_2 \\
\end{pmatrix}
\]

(4)

A model with a constant value (normally the mean) is written as

\[
G = \begin{pmatrix}
1 \\
\vdots \\
1 \\
\end{pmatrix}, \quad A = \alpha_1
\]

(5)

As we can separate the basis functions from the coefficients, this is a linear problem (linear in terms of the unknown coefficients). Here, we use only constant value regression functions between changepoints. However, the approach we present generalises to any linear function, allowing for more gradual transitions between states (e.g., linear drift).

2.1.2 Data likelihood
As commonly assumed in many geochemical problems, we make an implicit assumption that the noise associated with the data is normally distributed with a mean of $\bar{\varepsilon}$ and a variance, $\sigma^2$, i.e. if we have individual errors represented by $\varepsilon_i$, then we can write the distribution of the errors as

$$p(\varepsilon_i) = N(\bar{\varepsilon}, \sigma^2)$$  \hspace{1cm} (6)$$

and the variance of this noise distribution may be poorly (or not) known. This is equivalent to the residuals between the observations and the predictions being normally distributed with mean of zero and a variance of $\sigma^2$. If we have reliable and representative estimates of the noise variances, we can use the values in the likelihood instead of $\sigma$. In practice, the noise term determines the uncertainty in fitting the data which, due to geological complexity, is usually greater than the reported analytical precision. Given this assumption, a Gaussian likelihood function for observed data (using $d = d_{\text{obs}}$) in the $i$-th partition, given the predictions with a particular set of model parameters ($m$) is

$$p(d_i|m) = \prod_{j=1}^{k_i} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\left[\frac{(d_{i,j} - f_i(x))^2}{\sigma^2}\right]}$$  \hspace{1cm} (7)$$

where the subscript $j$ refers to the $k_i$ data in the $i$-th partition (i.e. the region bounded by changepoints $c_i$ and $c_{i-1}$, with $c_0$ defined the lowest value of the data locations) and $d_{i,j}$ is the $j$-th observation in partition $i$.

More generally, for $n$ partitions and $N_d$ different datasets, we have the joint likelihood function,

$$p(d_1, d_2, \ldots, d_{N_d}|m) = \prod_{i=1}^{n} \prod_{i=1}^{k_i} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\left[\frac{(d_{i,j} - f_i(x))^2}{\sigma^2}\right]}$$  \hspace{1cm} (9)$$
Given we subsequently assume a constant value in a partition, an appropriate form to chose values for the predictive function \( f \) is a normal distribution centred on the mean value (and a variance equal to the variance) of the data in that partition. In this case, values for the regression model parameters \( A \) can be drawn from this distribution while the mean is the most probable value from the posterior distribution in a Bayesian formulation, equivalent to the maximum likelihood estimate. We could use a similar approach for any linear function with unknown coefficients, using standard least squares inverse methods (e.g. Menke 1989) to find the maximum likelihood values and the covariance matrix for the coefficients for a given partition, and draw samples for the parameters.

2.1.3 Prior distributions

In a Bayesian formulation, we need to specify prior distributions on all unknown parameters. The priors reflect what we consider reasonable to assume about the possible values for each parameter. Bayes’s rule lets us use the information about the model parameters contained in the data to update our prior information (i.e. to produce the posterior distribution). If the posterior distribution is the same as the prior, then the data have told us nothing we did not already know.

We use the law of hierarchical probability to write the priors as

\[
p(m) = p(n)p(c,A,\sigma | n) \\
= p(n)p(c | n)\prod_i^N p(A_i | c, n)p(\sigma_i | c, n)
\]

(10)

where \( N_d \) is the number of datasets. The choice of priors in our formulation is straightforward (usually uniform between a specified minimum and maximum value) and are given in Supplementary material, SM2.
3. Markov chain Monte Carlo method for solving the changepoint problem

In using the Bayesian formulation described above, our goal is to generate a collection, or ensemble, of values approximating the posterior distribution, whose form we do not know in advance. As we also do not know the number of changepoints, the problem becomes what is known as transdimensional, where the number of model parameters itself becomes an unknown. To solve this problem we use a generalised version of Markov chain Monte Carlo (MCMC) sampling, known as Reversible Jump MCMC (Green 1995, 2003). A general introduction to MCMC methods is given by Gilks et al. (1996), a review of transdimensional Markov chains is given by Sisson (2005) and Gallagher et al. (2009) present an overview of the general methodology and its application to Earth Science problems. Specific applications to Earth Science problems have been presented by Malinverno (2002), Malinverno and Leaney (2005), Stephenson et al (2006), Jasra et al. (2006), Sambridge et al. (2006), Hopcroft et al. (2009), Charvin et al. (2008), Bodin and Sambridge (2009) and Piana Agostinetti and Malinverno (2010). We give brief overview of MCMC in the Supplementary material, SM1, while here we describe the aspects important for the transdimensional changepoint problem. The mathematical details for the MCMC implementation we adopt for the changepoint problem are given in Supplementary material, SM2 and SM3.

MCMC is an iterative method, and at each iteration, we consider 2 sets of model parameters, the current and proposed models ($m_c$ and $m_p$). The procedure for a given iteration can be described as follows

(i) Randomly perturb the current model to produce the proposed model

(ii) Randomly accept or reject the proposed model (in terms of replacing the current model), according to the acceptance criterion ratio (see equation A1.2 in Supplementary material, SM1).
In principle, after many iterations, the MCMC sampler should converge to a stable configuration (that is sampling according to the posterior distribution) and the final stage of MCMC is to use the sampling to infer characteristics and uncertainties for the model.

(i) Model perturbations/moves

As stated above, the sampling should converge to the target posterior distribution. However, the efficiency of the method does depend on choosing a reasonable proposal function to avoid moving too slowly around the model space as a result of the perturbations being either too small or too big (see figure 2 of Gallagher et al. 2009). The scale of model parameter perturbations can be tuned to achieve a reasonable balance between accepting and rejecting proposed models.

For the changepoint problem, we define 5 types of model perturbation or move:

1. Change the location of a changepoint
2. Change the regression function (mean estimate) in a partition between 2 changepoints
3. Change the value of noise for a dataset (if appropriate for that dataset)
4. Add a new changepoint (birth)
5. Remove an existing changepoint (death)

Each type of move has specified probability of being selected (which forms the jump proposal, $R$, referred to in Supplementary Material), and these probabilities need to sum to unity. In our problem, these are set to 0.2, 0.15, 0.15, 0.25 and 0.25 in the order listed above. The birth and death probabilities need to be modified to avoid having more/less changepoints than the maximum/minimum values ($n_{\text{max}}/n_{\text{min}}$). We do this by setting the birth (death) probability to 0.5 (0.0) for a model with $n_{\text{min}}$ changepoints, and the death
(birth) probability to 0.5 (0.0) for a model with $n_{\text{max}}$ changepoints. Having selected a perturbation type for a particular iteration, all other parameters are kept constant.

(ii) Acceptance criterion

For the purposes of describing how we accept or reject the proposed model, we use a simplified form of acceptance criterion ratio appropriate for 2 models with the same number of model parameters (and the full expressions for the transdimensional case are given in the Supplementary material). This can be written as a ratio of probabilities, given as

$$\alpha(m_p, m_c) = \min\left[1, \frac{p(m_p)}{p(m_c)} \frac{p(d|m_p)}{p(d|m_c)} \right]$$

where $\min[1, Z]$ means we take the minimum of 1 and $Z$. The terms $p(m)$ and $p(m|d)$ are the prior and likelihood probabilities for a particular model (and so define the posterior probability, at least up to the constant of proportionality). We have already introduced the concepts of the prior and likelihood functions (and where appropriate, the problem specific details are given in Supplementary material, 2).

The proposal probability, $q(m_p|m_c)$ determines how we move from the current model to proposed model (step (i) above). The theory underlying MCMC requires us to be able to reverse such a move (so we need to also include the reverse proposal probability in the ratio). If we consider prior distributions to be uniform (i.e. all models have the same probability), and if we have proposal distributions that are symmetric, then we can write equation 14 above as

$$\alpha(m_p, m_c) = \min\left[1, \frac{p(d|m_p)}{p(d|m_c)} \right]$$

This is just a ratio of the likelihoods (i.e. the probability of the proposed and current models producing the observed data). Thus if the proposed model fits the observed data better than
the current model (so it has a higher likelihood), then the likelihood ratio is > 1, and then
\( \alpha(.) \) is set to 1. Conversely, if the current model fits the data better than the proposed
model, the ratio is < 1, and then \( \alpha(.) \) is set to \( \frac{p(d|m_p)}{p(d|m_c)} \), which itself is always > 0.
The final step in an iteration requires us to generate a uniform random number, u, between
0 and 1, and compare this to \( \alpha(.) \). If \( u \leq \alpha(.) \) we accept the proposed model (and this
becomes the current model for the next iteration), otherwise we reject it (and we retain the
current model for the next iteration). From this, we can see that (given the assumptions
about flat priors and symmetrical proposal functions) we will always accept a proposed
model fits the data better as u is always \( \leq 1 \). If we consider a proposed model that fits the
data almost as well as the current model (say \( \alpha(.) = 0.95 \)), then, on average, we will accept
the proposed model 95% of the time. For a proposed model considerably worse than the
current model in terms of data fit (say \( \alpha(.) = 0.05 \)), then, on average, we will only accept
the proposed model 5% of the time. If the posterior distribution resembles a normal
distribution, we can see that this process will tend to concentrate the sampling under and
around the peak (the higher probability region), but also allows us to sample less good
models (out in the tails of the distribution). In fact, the number of accepted samples for
each model is proportion to the posterior probability of that model and the ensemble of
accepted models then is a good approximation of the posterior distribution.

For the first 3 moves described in (i) above, the number of model parameters is
constant and the acceptance ratio is given by equation 11. For the birth and death moves,
the dimensions of the current and proposed model are different and it is necessary to use
the acceptance ratio given in Supplementary Material 1 (equation A1.2). Intuitively, we
might expect that models with more parameters will tend to provide a better fit to the
observed data, and then that the sampler would tend always to increase the number of
model parameters towards the maximum. However, to demonstrate how MCMC operates
during transdimensional moves, we can consider again a simplified form of the acceptance
criterion as

\[
\alpha(m_p, m_c) = \min \left[ 1, \frac{p(m_p)p(d|m_p)}{p(m_c)p(d|m_c)} \right]
\]

(13)

If all the parameters are independent and we consider a proposed model with 1 more
parameter more than the current model (all other parameters being the same), then, with the
prior on the extra parameter given as \( p(m_{p,n+1}) \) we can write this as

\[
\alpha(m_{p,n}, m_c) = \min \left[ 1, \frac{p(m_{p,n+1})p(m_c)p(d|m_p)}{p(m_c)p(d|m_c)} \right]
\]

(14)

If both models fit the observations equally well (the likelihoods have the same value), then
as \( p(m_{p,n+1}) < 1 \), we have

\[
\alpha(m_{p,n}, m_c) = p(m_{p,n+1})
\]

(15)

Thus, in this special case, the acceptance probability is equal to prior probability on the
extra model parameter. In other words, even when the fit to the observations is as good as
the current model, the proposed model (with more parameters) is less likely to be accepted,
by a factor equal to the prior probability of the extra parameter.

More specifically, when we propose an increase in the number of changepoints
(birth), an increase in likelihood function will tend to encourage acceptance of the proposed
model. However the decrease in the prior ratio will tend to discourage acceptance due to
the increased dimensionality of the space. Overall, the algorithm always prefers a large partition rather than two small partitions with similar mean values (which would have similar likelihood values). This is an example of a property of Bayesian inference referred to as ‘natural parsimony’, which means that given a choice between a simple and complex models that provide similar fits to data, the simpler one will be favoured (e.g. Jeffreys and Burger 1992, Mackay 1992, Bretthorst 1993, O’ Ruanaidh and Fitzgerald 1996, Sivia 1996, Jaynes 2003).

(iii) Calculation of model and uncertainties

Typically the MCMC sampling is run for many (10^4-10^6) iterations, and includes an initial period during which the samples are not yet from the target posterior distribution. This is known as burn-in and these samples are discarded before making inference from the posterior distribution. Gilks et al. (1996) show examples of these characteristic behaviours as a guide for their recognition and we discuss this later with the examples. The post-burn-in samples should then provide a good approximation to the posterior distribution for the model parameters, i.e. p(m|d). This can be visualised for one model parameter by plotting a histogram. We can also calculate the expected model as a average, i.e.

\[
\overline{m} = \frac{1}{N} \sum_{i=1}^{N} m_i
\]  

(16)

which is effectively a weighted mean, in which the weighting is the posterior probability for each model. Similarly the variance and co-variance of model parameters are given by standard formulae. Finally, we can readily calculate the 95% credible intervals by ordering the samples for a particular variable, and simply identifying the upper and lower 2.5% of the distribution as the 95% credible interval.
Rather than choosing the best data fitting model, which tends to be overly complex, our preferred final solution is given by the expected model (equation 16) with 95% credible intervals around the regression function parameters, and the distributions on the number and locations of changepoints. When a large number of models are added together, their partitions overlap so the average model is continuous and smooth. An advantage of this is that we can produce a model that contains the features common to the majority of sampled models, but also can be more complex (yet smoother) than any individually sampled model.

4. Examples of changepoint modelling

We first use synthetic data to demonstrate that we can recover the known signal and noise terms. In this example, we discuss how to assess whether the MCMC sampler has performed adequately. Subsequently, we apply the method to 2 sets of real geochemical data, from peat cores in northeast Australia (Kylander et al. 2007) and eastern Tibet (Large et al. 2009).

4.1 Synthetic data

The synthetic data are shown in figures 2a,b,c. We randomly selected 4 changepoints and different mean value functions in each partition to produce 150 irregularly distributed samples for each dataset and added noise with different levels to each. We used these data, assuming unknown noise variance, to infer the distributions on the number and locations of changepoints and the noise variance.

As stated earlier, we tune the proposal functions to achieve adequate sampling of the model parameters. Among the more common ways to assess if these input parameters are appropriate are to examine the rate of acceptance (typically around 30%, although 10-
60% may be adequate, e.g. Brooks et al., 2003), and also the behaviour of the likelihood or model parameters as a function of iteration (they should show no long term trend, and ideally resemble white noise). However, for birth and death we can not readily control the acceptance rates, which can be much lower (<5%) in this problem.

We choose a proposal function scale ($\theta$ in equation A2.6) that is proportional to the range (maximum and minimum values) of particular parameter. We use 0.2, 0.05 and 0.025 of the prior ranges for the regression function (here the mean), the changepoint locations and the noise, respectively. We make exploratory runs in which we monitor the acceptance rates over $10^4$-$10^5$ iterations, and adjust the scaling parameters accordingly. In this problem, if the acceptance rate is too high, the scaling parameters are too small, and vice-versa.

Having tuned the proposal functions, we run the chain for $5 \times 10^5$ iterations, with a burn-in of $2.5 \times 10^5$ iterations.

In figure 3, we show the log likelihood (data fit), the number of partitions and the sampling for the 3 noise parameters. During the early part of sampling (figure 3a,b), there is clearly structure in the chain. The initial log-likelihood is about $-1.2 \times 10^4$, but even over the initial 5000 iterations we see that the sampler quickly arrives in better regions of the model space, and the log-likelihood increases rapidly, even though the number of changepoints is decreasing. Over the same sampling period, the noise values have not equilibrated either, the blocky structure, indicative of relatively poor movement around the model space (or mixing). In contrast, the post-burn parts of the chain show that the sampling appears to have reached equilibrium. There are no significant trends in the sampling (they look like white noise), and the number of partitions is sampled between 4 and 7.

As described in section 3, we use the post-burn-in samples to calculate the expected (or average) changepoint structure, the 95% credible intervals around it, the probability of
having changepoints over the range of the samples and also the mean and distribution on
the noise values for each dataset. In figures 2d,e,f we show the 3 datasets, with the mean
estimated noise value as error bars, together with the expected changepoint structure, and in
figure 4 we show the distributions on the noise parameters. It is clear, by comparison with
figure 2, the changepoint structure has been recovered well, with no spurious features, and
also the mean noise is in good agreement with the original values. In figure 5 we show the
distribution on the inferred number of changepoints, demonstrating that the inference leads
to about 80% probability there are 4. This is conditional on all the model assumptions (a
finite number of discrete changes with constant mean values in each partition), although
these are appropriate in this example.

To demonstrate the influence of different datasets, we ran each dataset
independently, using the same parameters as the joint run and the results are shown in
figure 6. Again the main changepoint structure is recovered, although we see that some small
scale artifacts have been introduced for individual datasets. This leads to a slightly different
distribution on the number of changepoints, although all 3 datasets still have 4 as the most
probable, with the frequency of 5 changepoints, relative to 4, being higher than in the joint
model (25% compared with 98% dataset 1, 65% dataset 2, and 30% dataset 3). In practice, it
is unlikely that we would be primarily interested in the absolute number of changepoints,
but rather in where changes are inferred to occur. If we have multiple datasets, and expect
them to have the same changepoint structure, then we recommend modelling them together
for consistency in the changepoint structure.
4.2 Real data examples

In the real data examples described below, we follow a similar approach as described above. We monitored the acceptance rate on all variable parameters, and where relevant we adjusted the proposal function scales to achieve an acceptance rate of 20-40%.

4.2.1 Lynch’s Crater, Australia

The first real data example we consider is from Kylander et al (2007), who undertook geochemical analyses on a 13m section equivalent to ~50Kyr from a peatbog at Lynch’s crater, north-east Australia. They reported Pb isotope and Rare Earth Element data, used as proxies for climate change, and in particular for variations in air-transported mineral dust sources. Here we follow that paper, and consider the $^{206}\text{Pb}/^{207}\text{Pb}$ isotope ratios and the Europium anomaly, $(\text{Eu/Eu}^*)_{\text{PAAS}}$, which is a measure of Eu$^{2+}$ fractionation from Eu$^{3+}$ relative to the adjacent ions, Sm$^{3+}$ and Gd$^{3+}$ (Eu$^*$ being the geometric mean of these two).

We first use this example to demonstrate the inference of changepoints either representing the noise with the analytical errors or estimating the noise variance directly from data. In terms of analytical errors, those for the Pb isotope ratio were determined from a long-term series (nearly 2 years) measurements of the NBS 981 standard. In the absence of an equivalent estimate from repeat measurements, we assumed 10% of the observed value for the Eu anomaly. The mean noise variances on Pb and the Eu anomaly are $3.86 \times 10^{-4}$ and 0.123, respectively.

In figure 7 we show the inferred changepoint structure using these specified errors in the data likelihood, and also the case in which we also infer the noise variances, for each dataset in terms of a probability distribution (figure 8). Clearly, the structure in the first case is dominated by the Pb isotope ratios (which have relatively small analytical errors). The mean number of changepoints is 106 ($\pm 4$, 1σ) and it is difficult to make meaningful
sense of these results. In the second case, where we estimate the noise variance, the mean number of changepoints is 6 (±1, 1σ), and correspondingly, the mean of the noise variances on the Pb isotope data and the Eu anomaly are 0.00913 and 0.0341, respectively. In this case, the inferred noise variance on the Eu anomaly is about 3-4 times smaller than the assumed 10%, while for the Pb isotopes, it is about 24 times larger than the analytical level. In terms of the inferred changepoint locations, the major peaks around 150 and 620m correspond to the two most significant changes inferred by Kylander et al. (2007). The shallowest changepoint is related to a change to warm, wet conditions while the second is change from humid to arid. The 3rd peak around 820m in our results also corresponds to a lesser change inferred by the earlier work. We refer the reader to Kylander et al. (2007) and Muller et al. (2008), for a more detailed discussion of the environmental significance of these changes.

We also ran the 2 datasets individually, assuming the errors were unknown, and the results were very similar in terms of the error distributions. The Pb isotope data however only required one significant changepoint (around 820m), while the Eu anomaly data produced essentially the same result as the joint modeling.

4.2.2 Hongyuan, Eastern Tibet

Large et al. (2009) presented a series of geochemical and physical property measurements from a 6m deep section, equivalent to ~10Kyrs, of the Hongyuan peatbog in eastern Tibet. The aim of this study was to assess the relative influences of the Indian and east Asian monsoons, and to relate this to other inferences of climate variations in China. Here, we use the C, N, H and δ13C analyses, together with the bulk density and carbon density to make quantitative inference of changepoints. In this case, we have no specific information concerning the errors for each dataset, so we also need to infer the noise variance.
The changepoint and noise variance distributions are shown in figures 9 and 10. The noise levels are lower than the standard deviation of each dataset (~2-3 times lower), except for the $\delta^{13}C$ dataset, for which the inferred noise variance is similar in magnitude to the variation in data. The summary diagram of Large et al. (2009) (their figure 7) compares their data to previous studies, and in particular of inferred periods of cold, dry (permafrost) periods relative to warmer, wetter periods. Thus our inferred changepoints should correspond to times when these conditions switch. Apart from the relatively low amplitude probability changepoint inferred around 200 cm and the recent variations (<50cm, attributed to disturbance as a consequence of Yak grazing by Large et al, 2009), the changepoints agree well with those inferred by a qualitative comparison of regional datasets from China by Large et al. (2009) (see figure 9).

Although we do not show the results here, we also ran these 6 datasets individually. As we might now expect, the details of the changepoint location structure differs between each dataset. Also, the mean values of the estimated noise levels were lower (by between 10 and 60%) than for the joint model. This latter result is also not unexpected as the joint modelling tends to compromise (increase) the noise variance to accommodate common changepoints for multiple datasets. While there is clearly common information, it is not easy to identify reliably the changepoints by considering these datasets individually. Again we recommend joint modelling of multiple datasets if we anticipate a common changepoint structure for a particular problem.

5. Summary

Changepoints can be defined as abrupt changes in trends (such as the mean, gradient or any function) over depth or time. In this paper, we have presented a new approach to changepoint modelling, applicable to multiple datasets with common changepoint
locations, allowing for unknown noise variance in each dataset. The approach is based on Bayesian transdimensional Markov chain Monte Carlo and we estimate the changepoint structure in terms of distributions for number and location of changepoints, the regression function parameters and the noise variance on multiple datasets. Here, we have considered the regression function in terms of a constant value between 2 changepoints, but the approach generalises to any linear function of the data. In any transdimensional problem, the solution (i.e. the number and location of changepoints) is strongly influenced by the assumed noise variance. Our approach, in allowing us to estimate the noise variance directly from the data, is particularly useful when we do not have reliable estimates of the data error/noise, or perhaps only consider analytical errors (i.e. we neglect natural variation due to geological complexity) and so implicitly assume the data are more precise than is perhaps advisable. Furthermore, the Bayesian approach we adopt is naturally parsimonious and avoids inferring unwarranted complexity when finding the changepoint structure. Thus we expect to favour models with fewer changepoints, while still achieving an adequate fit to the observed data.

Using synthetic data, we have demonstrated that we can recover the changepoint structure and the noise variances reliably. When dealing with multiple datasets, we assume that all datasets contain the same changepoint locations, but the response, or regression functions, and noise variances are different. The approach we present can be generalised readily to allow for different noise variance between partitions, if required. Additionally, the different datasets can be irregularly spaced in depth (or time) and there is no need for the data to be sampled at the same depths (time). The details of the solutions will depend on which datasets are used (i.e. singly or jointly) and we recommend using joint modelling if the assumption of common changepoints is considered valid. This assumption is perhaps best assessed from the understanding of geochemical behaviour in different environmental
systems. Certainly, the results are more coherent and generally easier to interpret than by combining results from individual dataset modelling. Applications of the method to real datasets from NE Australia and eastern Tibet provide results in agreement with previous qualitative interpretations based on visual inspection. However, our approach is preferable as it is more objective, explicitly incorporates the noise variance (either known or unknown), allows us to assess quantitatively the relative importance of the inferred changepoint structure, and we obtain probability distributions on all parameters. Finally, directions for future work would be to consider transdimensional regression functions (for example we estimate the order of a polynomial which could be different between partitions) and to allow for uncertainty in depth to age conversions (which will be important when comparing records from different locations).
Acknowledgements

We thank the French Australian Science and Technology (FAST) program for support during this work. We thank 2 anonymous referees for comments on an earlier version.
REFERENCES


McDermott, F., 2004. Palaeo-climate reconstruction from stable isotope variations in


Figure Captions
Figure 1. An example of the changepoint problem. We have a set of noisy data (dots), with a common noise variance (the noise is Gaussian, and $1\sigma$ value is shown as the error bar in the bottom right). The underlying function from which the data were generated is shown by the solid line. The function is discontinuous, with 4 changepoints at $x = 2, 5, 6, \text{and} 8$. In a real problem, the model parameters are the number of changepoints ($n$), the locations of the changepoints ($x_i, i = 1, n$) and the values of the function in each region (in this case, this is just the mean value of the data between each changepoint).

Figure 2
(a,b,c). The 3 synthetic datasets (grey dots), with 4 common changepoints. The noise scale ($\sigma$ from equation 7) is given in the top left, with an error bar $\pm 1\sigma$ shown just below. The true regression function for each dataset is shown by the solid line.
(d,e,f) Changepoint structure inferred for the 3 synthetic dataset. The solid line is the inferred function (relative to the lefthand axis), and the lighter dashed lines represent the 95% credible intervals on this function. The continuous lines represent the probability of a changepoint (relative to the right hand axis). The error bars are drawn using the mean value of the noise variances for each data set (see figure 4).

Figure 3.
(a) Log-likelihood (LL) and the number of changepoints ($n$) for the initial 5000 iterations.
(b) The sampled values for the 3 noise terms over the initial 5000 iterations.
(c) As (a), but during the post-burn-in sampling
(d) As (b), but during the post-burn-in sampling
Figure 4. Inferred distributions of the noise variances for each data set. The true values are shown as the heavy vertical lines.

Figure 5. Inferred distribution on the number of changepoints for the data in figure 3.

Figure 6. As figure 2,d,e,f,, but each data set was modelled individually.

Figure 7. Inferred changepoint model for the data from Kylander et al. (2007). Their inferred climatic-related varitations are indicated by the grey bar bars at the base of each graph.

(a) Pb isotope data using the analytical errors for the noise term.
(b) Eu anomaly data, using 10% of the observed value as the noise term
(c) As (a), but we infer the noise variance (mean value shown as the error bars)
(d) As (b), but we infer the noise variance (mean value shown as the error bars)

Figure 8. Inferred distributions of the noise variance for the two datasets of Kylander et al (2007).

Figure 9. Inferred changepoint model for the data from Large et al. (2009). Their inferred climatic-related variations are indicated by the grey bar bars at the base of each graph.

(a) H, (b) N, (c) C, (d) Total Carbon, (e) $\delta^{13}$C, (f) density

Figure 10. Inferred distributions of the noise variance for the 6 datasets of Large et al (2009). The sequence of graphs is the same as in figure 9.