Approximate simulation-free Bayesian inference for multiple changepoint models with dependence within segments

Jason Wyse[1], Nial Friel[2] and Håvard Rue[3]

[1]University College London, London, UK

[2]University College Dublin, Belfield, Dublin 4, Ireland

[3] Norwegian University of Science and Technology, Trondheim, Norway

May 2011

Abstract

This paper proposes approaches for the analysis of multiple changepoint models when dependency in the data is modelled through a hierarchical Gaussian Markov random field. Integrated nested Laplace approximations are used to approximate data quantities, and an approximate filtering recursions approach is proposed for savings in compuational cost when detecting changepoints. All of these methods are simulation free. Analysis of real data demonstrates the usefulness of the approach in general. The new models which allow for data dependence are compared with conventional models where data within segments is assumed independent.

1 Introduction

There is a substantial volume of literature devoted to the estimation of multiple changepoint models. These models are used frequently in econometrics, signal processing and bioinformatics as well as other areas. The idea is that "time" ordered data (where time may be fictitious and only refers to some natural ordering of the data) is assumed to follow a statistical model which undergoes abrupt changes at some time points, termed the changepoints. The changepoints split the data into contiguous segments. The parametric model assumed for the data usually remains the same accross segments, but changes occur in its specification. For example, in the famous coal mining disasters data (Jarrett 1979), disasters are usually assumed to follow a Poisson distribution where the rate of this distribution undergoes abrupt changes at specific timepoints. Fearnhead (2006) discusses how to perform exact simulation from the posterior distribution of multiple changepoints for a specific class of models using recursive techniques based

on filtering distributions. The class of models considered assumes data is independent within a homogeneous segment and the prior taken on the unknown model parameters for that segment allows analytical evaluation of the marginal likelihood for that segment. The paper of Fearnhead (2006) proposes a very promising step forward for the analysis of multiple changepoint models, where the number of changepoints is not known beforehand. The methods developed there allow for efficient simulation of large samples of changepoints without resorting to MCMC.

An obstacle which may prevent wide applicability of the methods discussed in Fearnhead (2006), is the requirement that the assumed model must have a segment marginal likelihood which is analytically tractable. However, such a requirement can usually not be fulfilled by models which allow for data dependency within a segment, a desirable assumption in many situations. Dependency is possible across regimes in some cases (see Fearnhead & Liu (2010)), but the assumption of independent data still holds. The main aim of this paper is to provide a solution to these issues and open up the opportunity for more complex segment models which allow for temporal dependency between data points. This is achieved by hybridizing the methods in Fearnhead (2006) and recent methodology for the approximation of Gaussian Markov random field (GMRF) model quantities due to Rue, Martino & Chopin (2009) termed INLAs (integrated nested Laplace approximations).

The INLA methodology provides computationally efficient approximations to GMRF posteriors, which have been demonstrated to outperform MCMC in certain situations (Rue et al. 2009). An advantage to such approximations is that they avoid lengthly MCMC runs to fully explore the posterior support and they also avoid the need to demonstrate that these runs have converged. Another advantage is that the approximations may be used to estimate quantities such as the marginal likelihood of the data under a given GMRF model, the quantity which is of main interest here to overcome the requirement of an analytically tractable segment marginal likelihood in Fearnhead (2006).

The R-INLA package Rue et al. (2009) for R-2.11.1 may be used to do all of the aforementioned approximations for a range of GMRF hierarchical models. It aims to give an off-the-shelf tool for INLAs. Currently the package implements many exponential family models; Gaussian with identity-link; Poisson with log-link; Binomial with logit-link; for many different temporal GMRFs; random effects models; first order auto-regressive; first and second order random walk (neither of these lists are exhaustive). The package also implements spatial GMRFs in two and three dimensions and is currently still evolving with new additions on a regular basis. Use of this package avoids programming for specific models as it allows the selection of any observational data model and selection of the desired GMRF through a one line call to the R-INLA package. The R-INLA package is used for all the computations on hierarchical GMRF models in this paper.

The remainder of this paper is organised as follows. Section 2 gives a brief review of recursions for performing inference conditional on a particular number of changepoints as given in Fearnhead (2006). In Section 3 possible computational difficulties

are discussed and solutions for these are proposed. Sections 4 and 5 analyze real data examples; the coal-mining data is analyzed using a model with dependency and this is compared with the analysis of Fearnhead (2006); and Well-log data (Ó Ruanaidh & Fitzgerald 1996) is analyzed with a model that allows for dependency between adjacent data points, such that the dependency relation may change across segments. Section 6 explores the possibility of detecting changepoints under the assumption of a stochastic volatility model. The paper concludes with a discussion.

2 Changepoint models

Fearnhead (2006) gives a detailed account of how filtering recursions approaches may be applied in changepoint problems. Some of the models considered there used a Markov point process prior for the number and position of the changepoints. Wyse & Friel (2010) demonstrated that the posterior distribution may sometimes be sensitive to the choice of the parameters for the point process. In this paper, the focus will be on performing inference for the changepoint positions after estimating the most probable number of changepoints *a posteriori*, although it is noted that the methods also apply to the case of a point process prior. Denote k ordered changepoints by $\tau_1, \ldots, \tau_k$. The likelihood for the data $\mathbf{y}_{1:n}$, conditional on the k changepoints and the latent field $\mathbf{x}$, assuming segments are independent of one another is

$$\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{j=1}^{k+1} \pi(\mathbf{y}_{\tau_{j-1}:\tau_j}|\mathbf{x}_j, \boldsymbol{\theta}_j),$$

where $\tau_0 = 0, \tau_{k+1} = n$, $\mathbf{x}_j$ represents the part of the GMRF $\mathbf{x}$ which belongs to the j^{th} segment, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{k+1}^{\mathrm{T}})^{\mathrm{T}}$ are the segment hyperparameters. Independent priors are taken on the members of Θ and the changepoints given their number. The prior taken on changepoints is assumed to have the product form

$$\pi_k^{\mathrm{cp}}(\tau_1, \ldots, \tau_k) = \prod_{j=0}^{k} \pi_k^{\mathrm{cp}}(\tau_j|\tau_{j+1}).$$

where $\tau_0 = 0, \tau_{k+1} = n$. Note that this prior is conditional on a given number of changepoints, k. The idea is to introduce a prior on k and use the hierarchical form

$$\pi(k|\mathbf{y}) \propto \pi(\mathbf{y}|k)\pi(k) \tag{1}$$

to find the most likely number of changepoints. Using this, the most likely positions for the changepoints can then be found.

2.1 Recursively computing the posterior

Let $L_j^{(k)}(t) = \Pr(\mathbf{y}_{t:n}|\tau_j = t-1, k)$. Then $L_j^{(k)}(t)$ is the probability of the data from time point t onwards given the j^{th} changepoint is at time $t-1$ and there are k changepoints

in total, meaning that there are $k-j$ changepoints between times t and n. It is possible to compute $L_j^{(k)}(t)$ in a backward recursion;

$$L_j^{(k)}(t) = \sum_{s=t}^{n-k+j} P(t,s) L_{j+1}^{(k)}(s+1) \pi_k^{\mathrm{cp}}(\tau_j = t-1 | \tau_{j+1} = s)$$

with j going from k to 1 and t going from $n-k+j-1$ to $j+1$, where $P(t,s) = \pi(\mathbf{y}_{t:s})$ is the marginal likelihood of the segment $\mathbf{y}_{t:s}$. The marginal likelihood of $\mathbf{y}_{1:n} (= \mathbf{y})$ under a k changepoint model may be computed as

$$\Pr(\mathbf{y}_{1:n}|k) = \sum_{s=1}^{n} P(1,s) L_1^{(k)}(s+1) \pi_k^{\mathrm{cp}}(\tau_1 = s). \tag{2}$$

2.2 Choice of changepoint prior and computational cost

It will be necessary to compute $\pi(\mathbf{y}_{1:n}|k)$ for a range of values, say $k = 0, \ldots, K$ in order to do inference for k using (1). This requires computational effort in $O(n^2K^2)$ and storage requirements in $O(nK^2)$ which could be costly. Both of these may be reduced by choosing an appropriate changepoint prior. One such prior, as used and noted by Fearnhead (2006), is to take changepoint positions distributed as the even numbered order statistics of $2k+1$ uniform draws from the set $\{1, \ldots, n-1\}$ without replacement. Doing this gives

$$\pi_k^{\mathrm{cp}}(\tau_1, \ldots, \tau_k) = \frac{1}{Z_k} \prod_{j=0}^{k} \delta(\tau_j | \tau_{j+1})$$

where $\delta(s|t) = t - s - 1$ and the normalizing constant $Z_k = \binom{n-1}{2k+1}$. Using this prior restricts the dependence of the prior on the number of changepoints to the normalizing constant only, meaning that

$$\begin{aligned}
L_{j+r}^{(k+r)}(t) &= \sum_{s=t}^{n-[k+r-(j+r)]} P(t,s) L_{j+r+1}^{(k+r)}(s+1) \delta(\tau_{j+r} = t-1 | \tau_{j+r+1} = s) \\
&= \sum_{s=t}^{n-k+j} P(t,s) L_{j+r+1}^{(k+r)}(s+1) \times (s-t) \\
&= \sum_{s=t}^{n-k+j} P(t,s) L_{j+1}^{(k)}(s+1) \times (s-t) = L_j^{(k)}(t).
\end{aligned}$$

Reusing these values gives a reduction by a factor of K in computational effort and storage requirements. The recursions are now

$$L_j^{(k)}(t) = \sum_{s=t}^{n-k+j} P(t,s) L_{j+1}^{(k)}(s+1) \delta(\tau_j = t-1 | \tau_{j+1} = s) \tag{3}$$

and

$$\Pr(\mathbf{y}_{1:n}|k) = \sum_{s=1}^{n} P(1,s) L_1^{(k)}(s+1)\delta(\tau_0 = 0|\tau_1 = s). \tag{4}$$

Then (4) is divided by Z_k to correctly normalize the prior and (1) is obtained by multiplying this by the prior weight for k changepoints $\pi(k)$. This prior will be used in the examples later.

2.3 Posterior of any changepoint

Since the prior on changepoints makes the changepoint model factorizable, it is possible to write down the posterior distribution of τ_j conditional on τ_{j-1} and k;

$$\Pr(\tau_j|\tau_{j-1}, \mathbf{y}_{1:n}, k) \propto P(\tau_{j-1}+1, \tau_j) L_j^{(k)}(\tau_j+1)\delta(\tau_{j-1}|\tau_j)/L_{j-1}^{(k)}(\tau_{j-1}+1).$$

This is used for the forward simulation of changepoints once the backward recursions have been computed. It is also used to give the modal changepoint configuration as in the examples later.

3 Approximate changepoint inference using INLAs

The essential ingredient of the approach presented in this paper is to replace the segment marginal likelihood $P(t,s)$ in the recursions

$$L_j^{(k)}(t) = \sum_{s=t}^{n-k+j} P(t,s) L_{j+1}^{(k)}(s+1)\delta(\tau_j = t-1|\tau_{j+1} = s)$$

with a segment marginal likelihood approximated using INLA. It is the case that $P(t,s)$ needs to be available in closed form to use a filtering recursions approach. This will never be the case for hierarchical GMRF models, which can account for within segment dependency. However, INLAs can be used to get a good approximation to $P(t,s)$ for hierarchical GMRF segment models. This opens up the opportunity for more realistic data models in many cases. There are also two other advantages: the posterior of the number of changepoints may be well approximated for model selection; and the posterior of any given changepoint can be computed to a high degree of accuracy.

There are two potential drawbacks of the proposed approach however. The first is that it could require fitting a GMRF model to a very small amount of data, which could be limiting depending on the complexity of the within-regime model. For example, at least five data points would be required to make fitting a first order auto-regressive random field feasible. This means that for the approach to be reasonable it may be necessary to expect changepoints to be quite well separated. The second potential drawback contrasts with the first. For large amounts of data, using INLAs to compute the $n(n+1)/2$ segment marginal likelihoods necessary to compute the recursions (3)

could be costly. The next section proposes a way to overcome both of these problems simultaneously, while still retaining almost all of the advantages of using a filtering recursions approach. This proposed solution is termed reduced filtering recursions for changepoints (RFRs).

3.1 Reduced filtering recursions for changepoints

The main idea of RFRs is to compute the recursions at a reduced number of time points and approximate the full recursions (3). The recursion is not computed at every time point which takes $O(n^2)$ computation. The motivation is that if segments have a reasonable duration, changepoints can be detected in the region where they have occurred.

An analysis using RFRs permits a changepoint to occur at some point in the reduced time index set $\{t_1, \ldots, t_N\}$ with $t_i < t_j$ for all $i < j$. The assumption is that if there is a changepoint between t_i and t_{i+2} it can be detected at t_{i+1}. For convenience, define $t_0 = 0$ and $t_{N+1} = n$. The spacing of the t_i is an important issue. If the spacing is too wide, then changepoints will not be detected. If the spacing is too narrow, many points are required to represent the entire data, thus increasing the computation time. If there is little prior knowledge of where changepoints occur the natural choice is equal spacing; $t_i = ig$ for some choice of g. The following example briefly explores the choice of g and makes the preceding discussion clearer.

Consider the data simulated from a Gaussian changepoint model shown at the top of Figure 1(a) with a clear change at 97. Searching for one changepoint, the bottom three plots in Figure 1(a) show the posterior probability of a changepoint for reduced time index sets given by $g = 1, 5, 10$. Note that $g = 1$ corresponds to the original recursions (3). For $g = 5$ the changepoint is detected at 95 and $g = 10$ detects it at 100. In both cases the changepoint is identified as the closest possible point to its actual position. Figure 1(b) shows a similar example, where this time one of the segments is very short (only 13 points). Again, the changepoint is identified at the closest possible position in the cases of $g = 1, 5$. In the case of $g = 10$ it is the second closest, possibly due to the noise in the data contaminating the separation of the two regimes. The magnitude of the regime changes in these examples are large for illustration. If the magnitude of the change is small it will be necessary to have a longer segment to identify its presence with high power. It is also the case that if changes are short-lived, a large value of g may cause changepoints to be missed.

3.1.1 Recursions on the reduced time index set

The changepoints are $\tau_1, \ldots, \tau_k$. The reduced time index set is $\{t_1, \ldots, t_N\}$. The changepoint prior is now defined on the set of numbers $\{1, \ldots, N\}$ and we let $c_j = r$ if $\tau_j = t_r$. That is, c_j corresponds to the changepoint position if time is indexed by $\{1, \ldots, N\}$ whereas τ_j gives the changepoint position in the reduced time index set $\{t_1, \ldots, t_N\}$. Define

$$R_j^{(k)}(r) = \Pr(\mathbf{y}_{t_r+1:n} | \tau_j = t_r, k).$$

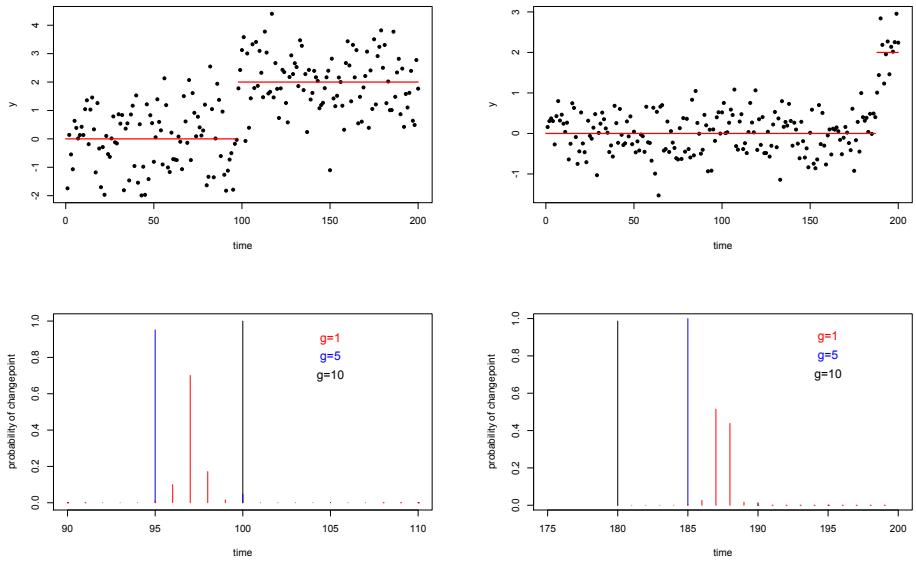

Figure 1: Results when searching for one changepoint in simulated Gaussian data for $g = 1, 5, 10$. It can be seen that the changepoint is detected at one of its closest neighbouring points in the reduced time index set.

For $r = N, \ldots, k+1$

$$R_k^{(k)}(r) = P(t_r + 1, n)\delta(c_k = r | c_{k+1} = N + 1).$$

Then recursively, for $j = k-1, \ldots, 1$ and $r = N-k+j-1, \ldots, j+1$

$$R_j^{(k)}(r) = \sum_{s=r+1}^{N-k+j} P(t_r + 1, t_s) R_{j+1}^{(k)}(s)\delta(c_j = r | c_{j+1} = s).$$

After computing these, the approximate marginal likelihood of the data conditional on k changepoints follows as,

$$\Pr(\mathbf{y}_{1:n}|k) \approx \sum_{s=1}^{N-k} P(1, t_s) R_1^{(k)}(s)\delta(c_0 = 0 | c_1 = s)/Z_k.$$

When the grid spacing g is not too large, that is n is greater than a reasonable multiple of g, the approximation to the marginal probability of k changepoints should be reasonable for the competing models. There are many computational savings with this approach. Using the RFRs decreases the number of marginal likelihood evaluations required to $n_r(n_r+1)/2$ where

$$n_r = \lfloor n/g + 1 - \mathrm{I}(g = 1) \rfloor.$$

3.1.2 Distribution of any changepoint

When the maximum *a posteriori* number of changepoints has been found, it is determined where the changepoints are most likely to occur on the reduced time index set. The distribution of c_j is

$$\Pr(c_j|c_{j-1}, \mathbf{y}_{1:n}, k) \propto P(t_{c_{j-1}} + 1, t_{c_j}) R_j^{(k)}(c_j) \delta(c_j|c_{j+1}) / R_{j-1}^{(k)}(c_{j-1}). \tag{5}$$

Instead of generating samples of changepoints, our focus is to deterministically search for the most probable changepoint positions *a posteriori*. The first changepoint detected on the reduced time index set will be

$$\hat{c}_1 = \arg\max_{c_1} \Pr(c_1|c_0 = 0, \mathbf{y}_{1:n}, k).$$

Conditioning on $\hat{c}_1$ the search proceeds for $c_2, \ldots, c_k$ in the same way. In general,

$$\hat{c}_j = \arg\max_{c_j} \Pr(c_j|\hat{c}_{j-1}, \mathbf{y}_{1:n}, k).$$

This procedure is repeated until the k changepoints $t_{\hat{c}_1}, t_{\hat{c}_2}, \ldots, t_{\hat{c}_k}$ are found.

3.1.3 Refining changepoint detection

After detecting changepoints on the reduced time index set, it is possible to refine the search and hone in on the most likely position of the changepoint. To begin, the changepoints obtained from the search above, $\tau_1^{(0)}, \ldots, \tau_k^{(0)}$ where $\tau_j^{(0)} = t_{\hat{c}_j}$, will all be multiples of g. Condition on the value of $\tau_2^{(0)}$ to update τ_1. Compute

$$P(1, \tau) P(\tau + 1, \tau_2^{(0)})$$

using INLAs for $\tau \in \{\tau_1^{(0)} - g + 1, \ldots, \tau_1^{(0)} + g - 1\}$. Then take $\tau_1^{(1)}$ to be the τ which maximizes this. Similarly $\tau = \tau_j^{(1)}$ maximizes

$$P(\tau_{j-1}^{(1)}, \tau) P(\tau + 1, \tau_{j+1}^{(0)}).$$

This procedure can be carried out just once, or repeated until there is no difference between updates.

This step does of course require additional computation. It may not be necessary in all cases to carry out a refined search. For example, the case of large n and small g would mean that refining the search would probably give little additional information. This approach should give near the global MAP for the changepoint positions. To ensure the global MAP is found it would be necessary to use some sort of Viterbi algorithm which would also use the approximated marginal likelihood values.

3.1.4 Simulation of changepoint positions

The approximate methods discussed here are entirely simulation free. It may be useful to allow for simulation of changepoints from their joint posterior, once all of the marginal likelihoods have been approximated as in Fearnhead (2006). Introduction of the RFRs makes this a little more difficult here. We expect that for larger values of g the distribution of the changepoints on the reduced time index set will be quite degenerate. This is since for a changepoint in a given region, it will always be detected at the same point in the reduced time index set. However, for smaller values of g it may still be possible to simulate changepoints on the reduced time index set and refine their positions by simulating from a distribution which conditions on the two neighbouring changepoints- a stochastic version of the approach to refine the position of the changepoint discussed above (Section 3.1.3).

3.1.5 Exploring approximation error and computational savings in a DNA segmentation example

To get a rough idea of the approximation error and the possible computational savings to be made by using RFRs, the methods were applied in a DNA segmentation task with a conditional independence model. This deviates from the general theme of the paper (to fit models relaxing conditional independence), however, it is included to offer some insight into RFRs in general.

DNA sequence data is a string of the letters A,C,G and T representing the four nucleic acids, adenine, cytosine, guanine and thymine. Interest focuses on segmenting the sequence into contiguous segments characterized by their C+G content. It is assumed that within a segment the frequency of constituent acids follows a multinomial distribution, so that

$$\pi(\mathbf{y}_{t:s}|\boldsymbol{\theta}) = \prod_{i=t}^{s} \theta_{\mathrm{A}}^{\mathrm{I}(y_i=\mathrm{A})}\theta_{\mathrm{C}}^{\mathrm{I}(y_i=\mathrm{C})}\theta_{\mathrm{G}}^{\mathrm{I}(y_i=\mathrm{G})}\theta_{\mathrm{T}}^{\mathrm{I}(y_i=\mathrm{T})}.$$

With a Dirichlet$(\alpha,\alpha,\alpha,\alpha)$ prior on $\boldsymbol{\theta}_{(t:s)}$ the marginal likelihood for a segment is

$$P(t,s) = \frac{\Gamma\{4\alpha\}}{\Gamma\{\alpha\}^4\Gamma\{s-t+1+4\alpha\}} \prod_{j\in\{\mathrm{A,C,G,T}\}} \Gamma\left\{n_j^{(t:s)}+\alpha\right\}$$

where $n_j^{(t:s)}$ is the number of occurences of acid $j \in \{\mathrm{A,C,G,T}\}$ in the segment from t to s inclusive.

The data analyzed is the genome of a parasite of the intestinal bacterium *Escherichia coli*. The sequence consits of 48,502 base pairs, and so will provide a good measure of the computational savings to be made for larger datasets when using RFRs. This data has previously been analyzed by Boys & Henderson (2004), who implemented a hidden Markov model using RJMCMC to select the Markov order. Here however, a changepoint model assuming data in segments are independent is applied. Cumulative counts of the nucleic acids over location along the genome are shown in Figure 2.

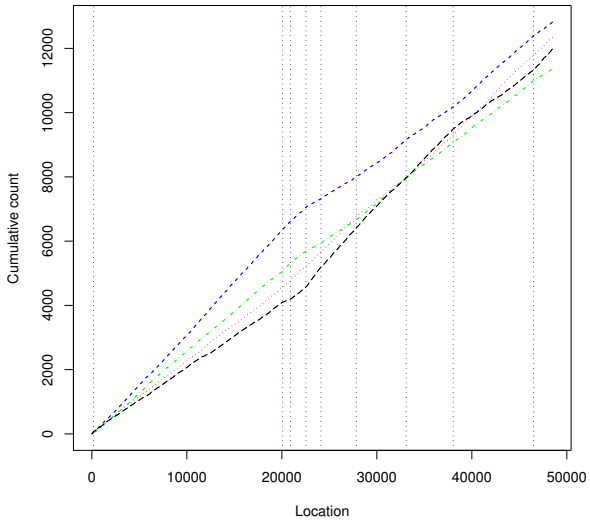

Figure 2: Cumulative counts of A,C,G,T for the DNA data. Identified changepoints are overlain (vertical lines).

g	1	5	10	15	20	25
Time taken (s)	1353.69	51.78	13.05	5.99	4.09	3.03
Changepoints	176	176	176	176	176	176
	20092	20101	20092	20092	20092	20092
	20920	20920	20920	20920	20920	20920
	22546	22585	22546	22546	22546	22546
	24119	24119	24119	24119	24119	24119
	27831	27831	27831	27831	27831	27831
	31226	31226	31226	31226	31226	–
	33101	33101	33101	33101	33101	33089
	38036	38036	38049	38011	38036	38036
	46536	46536	46536	46536	46501	46501

Table 1: Location of changepoints and computing time for DNA segementation example. As g increases there is little deviation in changepoint estimates. Reported changepoints are found after a refined search.

The RFRs were applied to this data using an equally spaced reduced time index set with $g = 1, 5, 10, 15, 20, 25$. The prior taken on the number of changes was uniform on $\{0, 1, \ldots, 20\}$. All runs were on a 2.66GHz processor written in C and the segment marginal likelihoods calculated in a step before the recursions were computed. Table 1 gives the identified changepoints and the computing time for each analysis. The value $g = 1$ corresponds to filtering recursions on the entire data. It can be seen that using RFRs does not appear to have a considerable effect on the detected changepoints. However, there are drastic differences in computing time- the RFRs for $g = 25$ give a 450 fold decrease in computing time with respect to recursions on the full data set. It can be seen that as the value of g increases there are slight deviations in the result. For example, with $g = 25$, nine changepoints is most probable *a posteriori*, compared with ten for all other g values. Also, for $g = 5$ the second and fourth changepoints are detected in a different position to all other values of g. This could be overcome by allowing for a wider search than just g points either side in the refined search (Section 3.1.3). Despite this, the computational savings are large, and the approach appears to successfully isolate the regions where changepoints occur in a large search space.

It should be noted that the computation of the marginal likelihoods can be nested, although this was not done here. For example, the marginal likelihood calculations for $g = 5$ could be reused for $g = 10, 15, \ldots$ and likewise, some of the calculations for $g = 10$ can be used for $g = 5$ if it is desired to perform analysis for different values of g.

3.2 Other approaches to save on computation

The RFRs approach reduces the computation necessary to perform analysis for changepoint models by introducing an approximation to full filtering recursions. This is complimentary with another approach to reduce computation suggested by Fearnhead (2006). There, recursions are "pruned" by truncating the calculation of the backward recursion when the terms to be truncated will only contribute negligibly to the overall value. This occurs in situations where it is clear that a changepoint will have occurred before a certain time in the future, and so considering times after this point does not lead to gains in information. These ideas could be used together with the approach presented here to gain extra speed up in computation. In particular, RFRs are useful for situations where segment sizes are large while pruning ideas are useful when the number of changepoints is large (so that calculations may be truncated often), so combining the two approachess should be useful for large scale problems with regular changes and will lead to even greater computational savings than just using RFRs alone.

4 Coal mining disasters

This data records the dates of serious coal-mining disasters between 1851 and 1962 and is a benchmark dataset for new changepoint approaches. It has been analyzed in Fearnhead (2006), Yang & Kuo (2001), Chib (1998), Green (1995), Carlin, Gelfand

& Smith (1992) and Raftery & Akman (1986), amongst others. In all of these analyses it is assumed that observations arise from a Poisson process. This Poisson process is assumed to have intensity which follows a step function with a known or unknown number of steps. These steps or "jumps" in intensity occur at the changepoints. Other models have also been fit to this data. For example, a smoothly changing log-linear function for the intensity of the Poisson process:

$$\lambda(t) = \nu \exp\{-\gamma t\}$$

(see for example Cox & Lewis (1966) and the original source of this data Jarrett (1979)). The log-linear intensity model would favour more gradual change, rather than the abrupt changes implied by changepoint models. There is an argument for some of the elements of such a model that allows for gradual change. Although, as noted in Raftery & Akman (1986), abrupt changes in this data are most likely due to changes in the coal mining industry at the time, such as trade unionization, the possibility of more subtle changes in rate could and should be entertained. A GMRF model applied to this data should be able to model gradual as well as abrupt change.

As in Fearnhead (2006) a week is the basic time unit. The data spans 5,853 weeks over 112 years. The latent field is taken as AR(1). This allows for an inhomogeneous Poisson process within segments, opening up the possibility for gradual change. The rate of the Poisson process is related to the field through a log-link function. More specifically,

$$y_i \sim \text{Poisson}(\lambda_i)$$

where

$$\lambda_i = \exp\{\alpha + x_i\}, \quad i = 1, \ldots, n.$$

The parameter α is an intercept and x_i follows an AR(1) process with persistence parameter ϕ.

Priors were chosen to loosely mimic the behaviour of the data. The priors chosen were

$$\begin{aligned} \sigma_{\mathbf{x}}^{-2} &\sim \text{Gamma}(4, 0.01) \\ \kappa &\sim \text{N}(3, 1.89^2) \\ \alpha &\sim \text{N}(0, 10^2). \end{aligned}$$

where we have reparametrized $\kappa = \text{logit}\left(\frac{1+\phi}{2}\right)$. Following Fearnhead (2006) and Green (1995), the prior on the number of changepoints was taken to be Poisson with mean 3.

A spacing of $g = 50$ was used. Figure 3 (a) shows the posterior distribution of the number of changepoints for the AR(1) latent field model. A two changepoint model is most likely, *a posteriori*. Figure 3 (b) shows the most likely position of these changepoints computed using the methods of Section 3.1.2. A plot of the log intensity of the Poisson process over the entire 5,853 weeks is shown in Figure 4, obtained by conditioning on the MAP changepoint positions from the two changepoint model. From this it can be argued that a model accounting for gradual changes in the rate of disasters

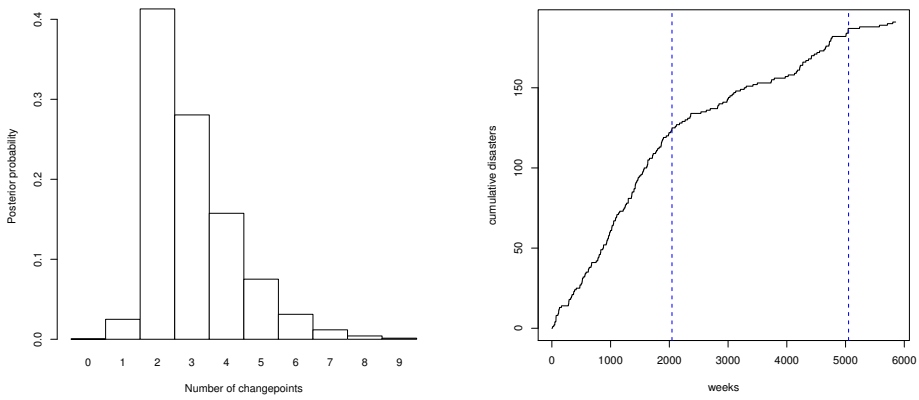

Figure 3: Coal mining data: results from an anlysis using INLAs and $g = 50$. The figure on the left shows the posterior distribution of the number changes while that on the right shows the cumulative counts of disasters and the changepoints indicated (blue dashed line).

is not entirely unjustified. There appears to be small fluctuations of rate around a mean rate. These fluctuations are treated differently to the two abrupt changes that are detected by the GMRF model.

There is a discrepancy between the posterior of the number of changepoints from RFRs given here and that given in Fearnhead (2006) (see Figure 1(a) there) which both allowed changepoints at all possible points in the data. This is a good opportunity to further investigate the approximation error introduced by using RFRs. Figure 5 shows the posterior number of changepoints obtained from using grids of size $g = 1, 5, 10, 15, 25, 50$ for the model and prior assumptions in Fearnhead (2006). It is clear that as the value of g increases, the RFRs become less sensitive to small or short lived changes for this model, as might be expected. However, at large values of g the ability to pick out two abrupt changes does not seem to diminish. As pointed out by one reviewer, a simple strategy for choosing g is possible by exploiting the nesting ideas outlined in Section 3.1.5. Starting out with a large value of g corresponding to a coarse search this may be gradually reduced to see how values of the approximated marginal likelihood for a given number of changepoints differs. Approximate marginal likelihoods computed for larger values of g may be recycled in doing computations for the more refined searches.

It is possible to compute approximate Bayes factors for the GMRF and independent data models conditional on there being a given number of changepoints. The marginal likelihood of the data conditional on k changepoints is approximately

$$\pi(\mathbf{y}_{1:n}|k) \approx \sum_{s=1}^{N-k} P(1, t_s) R_1^{(k)}(s) \delta(c_0 = 0 | c_1 = s) / Z_k.$$

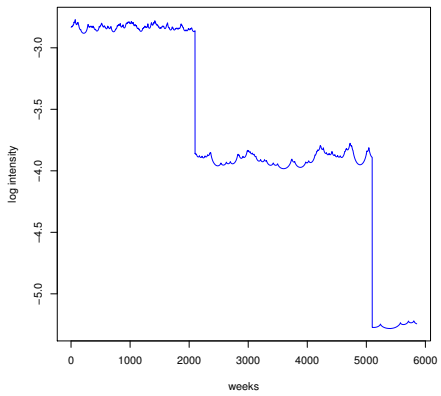

Figure 4: Coal mining data: Inferred log intensity by week.

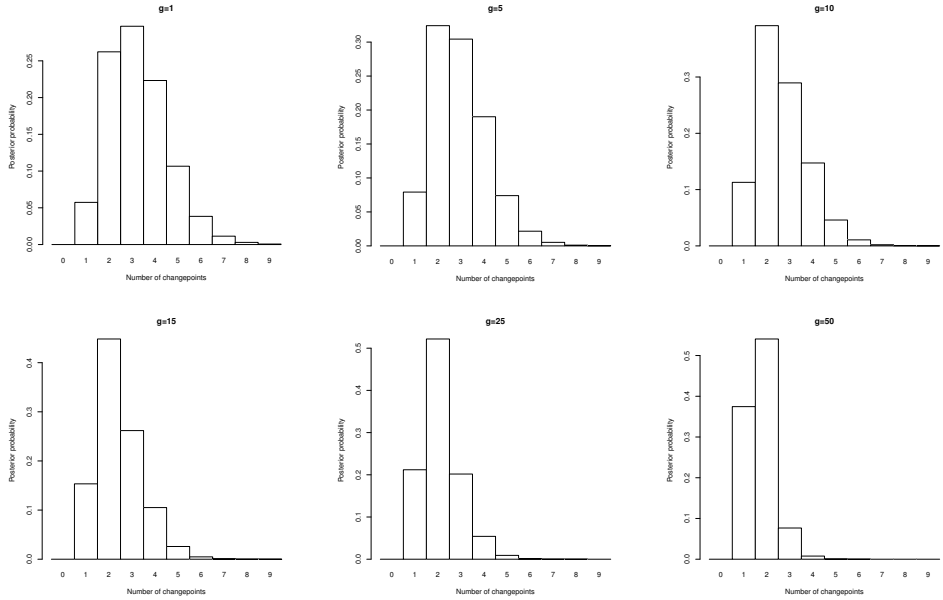

Figure 5: Investigating approximation error in RFRs; results from analyses of coal mining disasters with different values of g using the model from Fearnhead (2006).

The different models are characterized by model assumptions and consequently the way in which the segment marginal likelihoods are computed;

$$P_{\text{INLA}}(t,s) \quad \text{and} \quad P_{\text{ANALYTIC}}(t,s).$$

The approximate Bayes factor for the GMRF model versus the analytic model conditioning on k changepoints is given by

$$\mathcal{B}_k = \frac{\pi_{\text{INLA}}(\mathbf{y}|k)}{\pi_{\text{ANALYTIC}}(\mathbf{y}|k)}.$$

For a one changepoint model, this was $\mathcal{B}_1 = 4.63$ and for two changepoints it was $\mathcal{B}_2 = 5.25$. This implies that there is more support for the GMRF model in these cases, suggesting that modelling small scale variation in the rate of disasters is worthwhile. This supports the interpretation of Figure 4. It is well known that Bayes factors can be sensitive to prior assumptions. In this case prior hyperparameters have been chosen with care for both the independent data model and the GMRF model. A change in these choices could potentially lead to a different strength of conclusion as to which is the best model. However, it is still promising in this setting to see that modelling the dependency in the data appears worthwhile.

5 Well-log data

The Well-log data (Ó Ruanaidh & Fitzgerald 1996) records 4050 measurements on the magnetic response of underground rocks obtained from a probe lowered into a bore-hole in the Earth's surface. The data is shown in Figure 6. The model fitted here aims to account for dependency in the nuclear magnetic response as the probe is lowered into the bore-hole. This is an improvement on the independence model fitted in Section 4.2 of Fearnhead (2006); as the probe lowers, it moves through different rock strata and some will have greater depth than others. Therefore, it would be expected to see some correlation between observations arising from rock strata of the same type. Fitting this model can also reduce the detection of false signals as changepoints. See Fearnhead & Clifford (2003) for a discussion of the issue of outliers in Well-log data.

Since this is a large data set ($n = 4050$) a larger value of g should be used to isolate regions where changepoints occur. This vastly reduces the computational time required for the necessary approximations for data of this size. Analyses using $g = 10, 25, 50$ were carried out, choosing the prior parameters using the information obtained from an analysis using MCMC and an independent data model. In each instance numerical instability prevented the recursions on the reduced time index set from being computed. This happened because the scale of the data is so large ($\sim 10^5$). In general, measures need to be introduced to prevent numerical instabilities in these types of recursions. In the computations of the RFRs a measure similar to those in Fearnhead (2005) (changepoint models) and Scott (2002) (hidden Markov models) was employed.

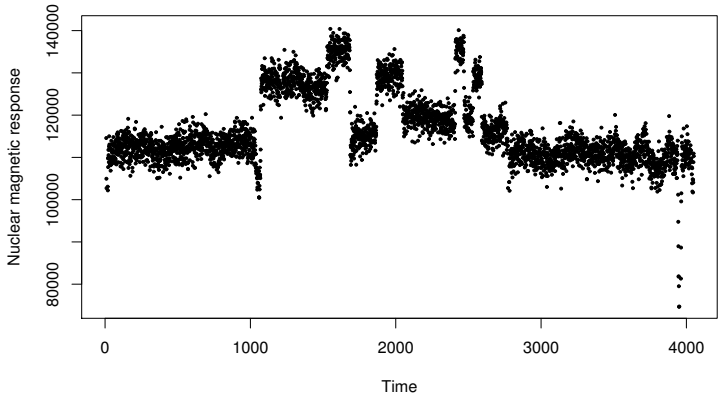

Figure 6: Well-log data. Observations are the nuclear magnetic response recorded by a probe being lowered into a bore-hole in the Earth's surface.

This consisted of two steps to ensure stability. Firstly, compute

$$\frac{R_j^{(k)}(r)}{R_{j-1}^{(k-1)}(r+1)} = \sum_{s=r+1}^{N-k+j} \delta(c_j = r | c_{j+1} = s) \exp\Big\{\log P(t_r+1, t_s) + \log R_{j+1}^{(k)}(s) - \log R_{j-1}^{(k-1)}(r+1)\Big\}$$

and then

$$\log R_j^{(k)}(r) = \log R_{j-1}^{(k-1)}(r+1) + \log\left(\frac{R_j^{(k)}(r)}{R_{j-1}^{(k-1)}(r+1)}\right).$$

The reason these do not work here is that the large scale of the data means that $\log P(t_r+1, t_s)$ is much larger than usual, since it is the marginal likelihood of $g = 10, 25, 50$ points. It thus makes the argument to the exponential function in the first stabilizing equation cause instabilities at some points. This then carries through the remainder of the recursions.

A simple way to overcome the issues is to just do an equivalent analysis of the data on a smaller scale, so that large $\log P(t_r+1, t_s)$ is avoided. Simply dividing the data by its sample standard deviation s reduces the scale appropriately. The parameters for the prior specification were also adjusted to allow for the difference in scale to give the priors

$$\begin{aligned} \sigma_{\mathbf{y}}^{-2} &\sim \text{Gamma}(1, 0.01) \\ \sigma_{\mathbf{x}}^{-2} &\sim \text{Gamma}(1, 0.01) \\ \kappa &\sim \text{N}(5, (\sqrt{10})^2). \end{aligned}$$

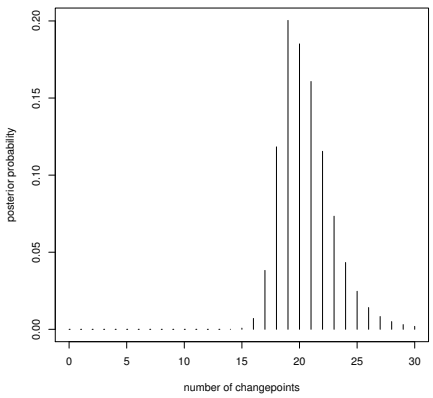

Figure 7: Posterior of the number of changepoints for the Well-log data fitting an AR(1) GMRF model. This suggests the most likely number of changepoints *a posteriori* is 19.

where $\kappa = \text{logit}\left(\frac{1+\phi}{2}\right)$. The prior on κ here gives most prior weight to values of ϕ in $[0.9, 1)$ (about 93%). This will allow the possibility for the AR(1) GMRF model to closely approximate the behaviour of a random walk of order one. However, it still allows the freedom for the dependence pattern to vary across segments. Fearnhead (2006) fits a random walk model of order one to this data, showing that a latent field can be robust to short lived changes and outliers for Well-log data. A uniform prior on $\{0, \ldots, 30\}$ was taken for the number of changepoints.

For the final analysis g was taken to be 25. This reduced the necessary number of approximate marginal likelihood approximations from roughly 8.2×10^6 (for $g = 1$) to 1.3×10^4; over 600 times less. The computations for these approximations took about a day of computing time. This appears lengthly, however this should be judged along with the fact that the model is more flexible and that the mean signal can be estimated at every point in the data. Figure 7 shows the posterior probability of the number of changepoints. The mode is at 19, but there appears to be support for up to 22. Conditioning on 19 changepoints, their locations were determined using the search strategy outlined in Section 3.1.2. These locations were then refined to hone in on the actual changepoint positions. Conditioning on these positions inference was carried out for the latent field. This is shown in the top figure of Figure 8. The field appears to follow the trend of the data closely, while the changepoint model caters for abrupt change. Fearnhead (2006) compared the results of a first order random walk field to those from an independent Gaussian model for the data. Similarly, the results from the GMRF model here are compared with those obtained using an MCMC sampler with an independent data model on the Well-log data. For comparison, the 54 most likely changepoints (mode of posterior) were taken from the independent Gaussian model, and segment means were computed conditional on these (bottom of Figure 8). It can be seen that the independent model is sensitive to changes in the mean and is conservative when

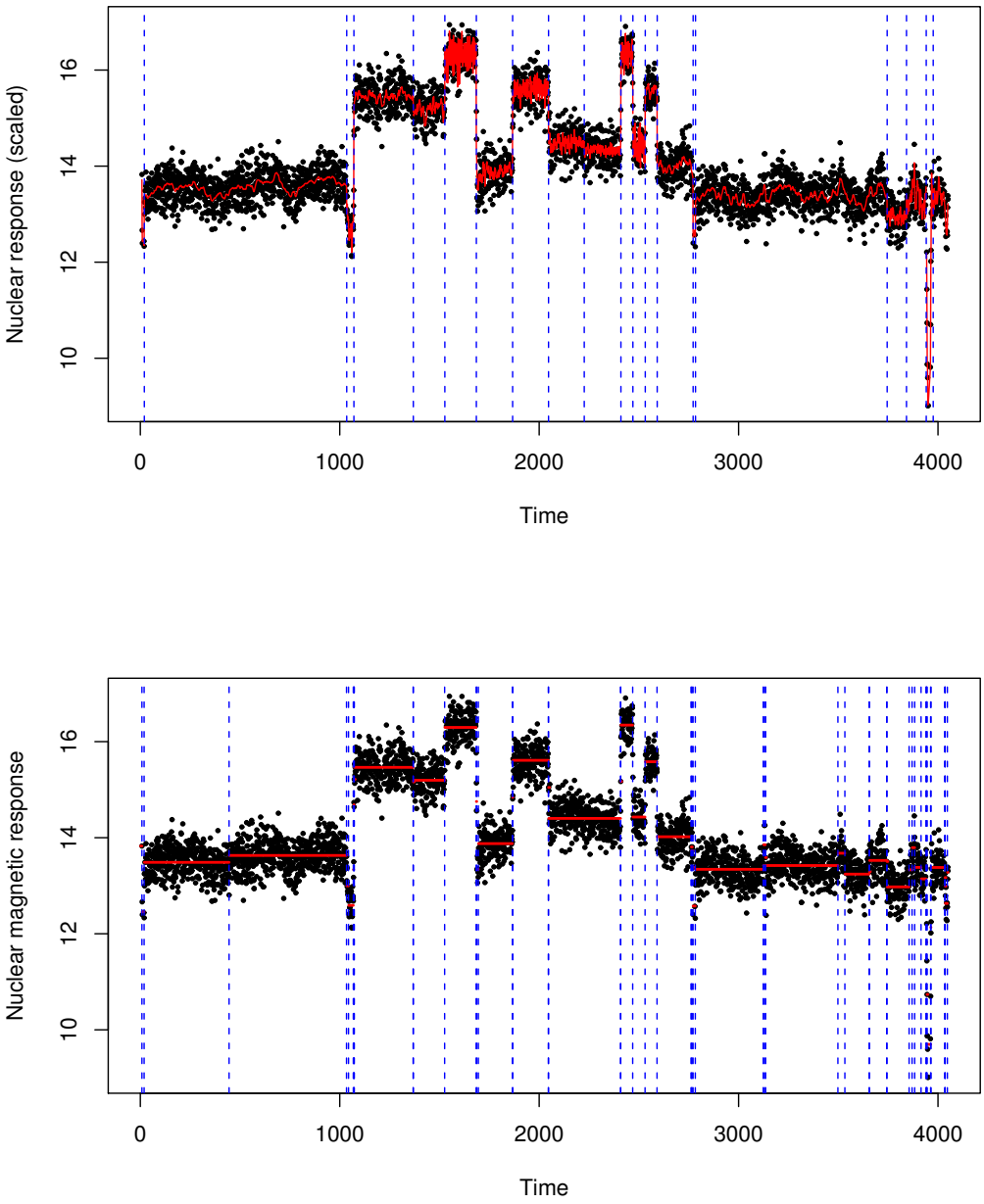

Figure 8: Well-log data: results from RFRs and INLA (top) and independent data model.

Segment	1	2	3	4	5	6	7	8	9
ϕ	0.9	0.8	0.9	0.7	0.8	0.9	0.8	0.9	0.8
$2\log\beta$	0	2	0	1	0	0.5	0	0.25	0
$\sigma_{\mathbf{x}}$	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01

Table 2: Segment parameters for simulated stochastic volatility data.

inferring changepoints (more rather than less). The GMRF model however appears to be more robust to noisy data points and only infers changepoints when abrupt changes occur in the field.

6 Stochastic volatility data

The aim of this section is to explore whether INLAs and RFRs can be used to estimate changepoint models where segment observations are assumed to arise from a stochastic volatility model. To this end, the approach proposed could potentially used to detect shocks in financial and other time series. The segment model assumed is

$$y_i \sim \mathrm{N}\left(0, \beta^2 e^{x_i}\right), \quad i = 1, \ldots, n,$$

with $\mathbf{x}$ following an AR(1) process with persistence parameter ϕ and innovation variance $\sigma^2_{\mathbf{x}}$ where $2\log\beta$ may be interpreted as an intercept for the volatilities. Data in different segments are assumed independent, so that concern here is only in the complex intra segment correlation structure.

The approach was applied to a simulated data set of length 1000 with eight changepoints at times 200, 400, 600, 700, 800, 850, 900 and 950, shown along with the corresponding log latent squared volatilities in Figure 9. Segment parameters were chosen as outlined in Table 2. The length of the segments were reduced as well as the magnitude of the regime change to see how powerful the approach is in detecting smaller and smaller changepoints.

The priors assumed for the analysis were

$$\begin{aligned} \sigma_{\mathbf{x}}^{-2} &\sim \mathrm{Gamma}(30, 0.02) \\ \kappa &\sim \mathrm{N}(3, 1) \end{aligned}$$

and the computations were done for a reduced time index set with spacing $g = 5$. Priors were chosen to loosely mimic the behaviour of the data. The prior chosen for the number of changepoints was Poisson(5).

Figure 10 shows the posterior of the number of changepoints with most support for six changes, but reasonable support for any number from five to eight. The changepoint positions found using the search strategy of Section 3.1.3 while conditioning on six changepoints were 192, 400, 595, 697, 806 and 939. These changepoints are shown with

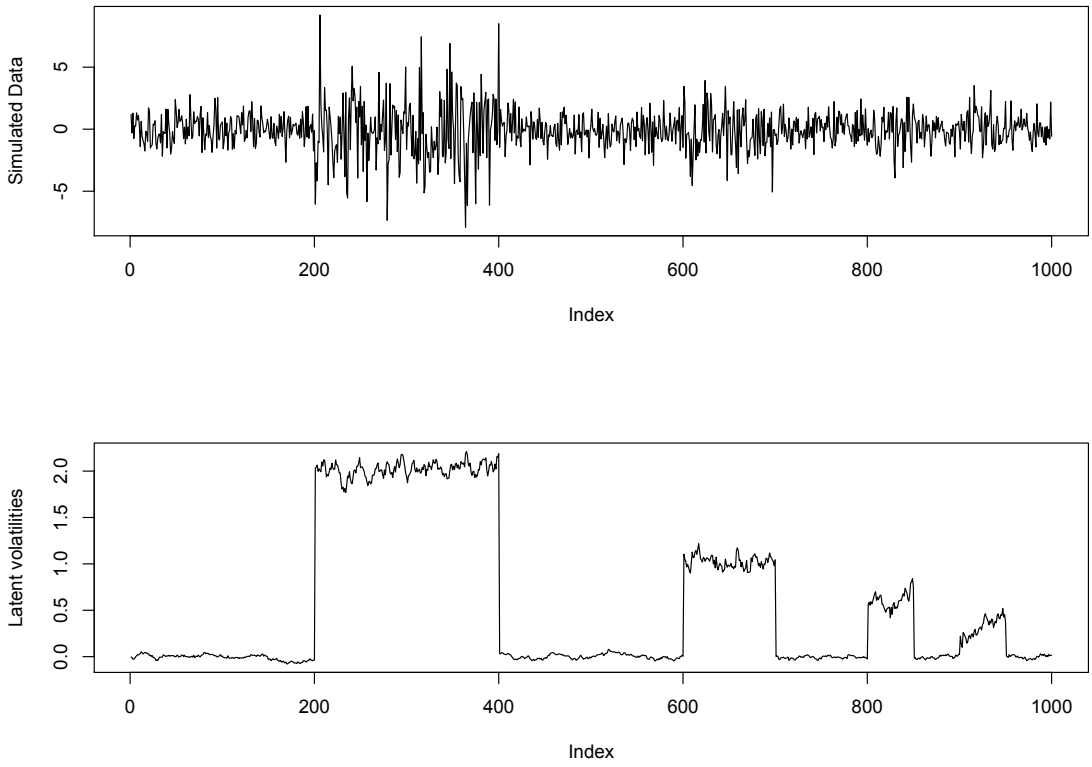

Figure 9: Simulated stochastic volatility data with the corresponding latent log squared volatilities shown on the bottom.

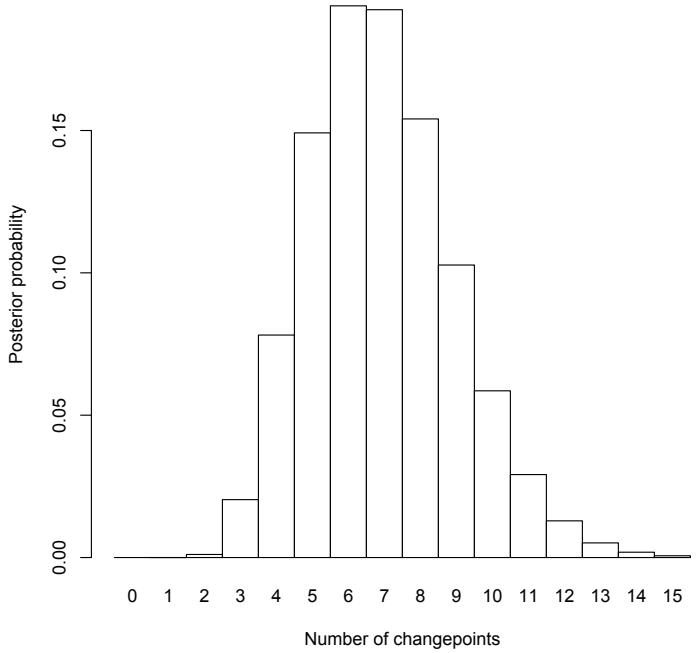

Figure 10: Posterior of the number of changepoints for simulated stochastic volatility data.

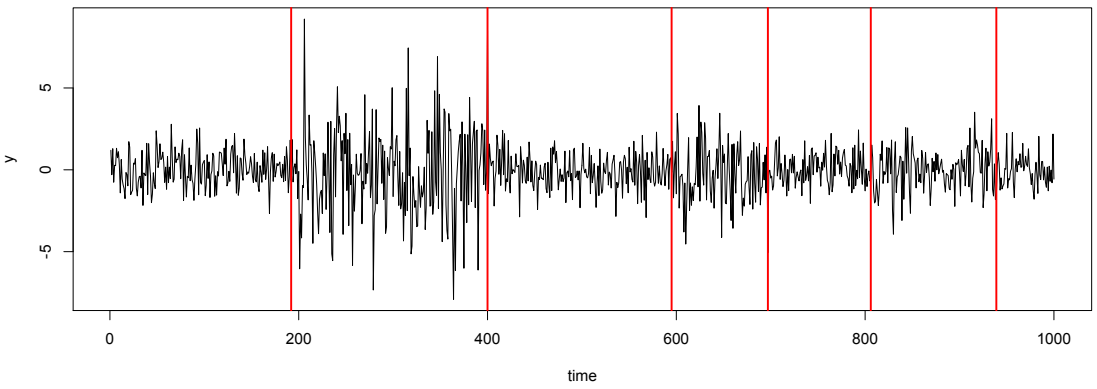

Figure 11: MAP changepoint positions conditioning on six changepoints.

the data in Figure 11. The changepoints are not detected at their exact positions, but are roughly close to the true positions given the scale of the data. The two changepoints at 850 and 900 are missed by this estimation strategy. These two changes are small in magnitude and are barely noticeable by simply looking at the data. In this situation it may be too difficult for any estimation strategy to distinguish between noise and a weak signal.

7 Discussion

This paper demonstrates two new useful approximate methods for changepoint problems when the assumption of independent data is relaxed. The first of these was INLAs, a new approximate inference method for GMRFs due to Rue et al. (2009). This allows the marginal likelihood for complex segment models to be evaluated approximately, so that it may be used for an approximate filtering recursions approach.

Some computational considerations led to the second proposed method. Instead of performing filtering recursions analysis on the entire data, RFRs were introduced so that recursions may be computed only on a reduced time index set, thus using all of the data, but only searching for changepoints in the general region where they occur. It was demonstrated that this method can be useful in cutting computation time for larger datasets by applying it to a DNA segmentation example with about 49,000 data points.

The hybrid INLAs-RFRs methodology was applied to three different data examples. The first of these was an analysis of the coal mining disasters data where the model allowed for small scale variation in the intensity of the process and allowed for week to week dependency. This new model was more supported by the data than the usual step function intensity models which are often fitted. This was demonstrated by approximate calculation of Bayes factors for the GMRF model and the independent data model for one and two changepoint models. The GMRF model out-performed the independent data model in both cases. The second example was an analysis the Well-log data of Ó Ruanaidh & Fitzgerald (1996). It was shown that allowing for segment dependency can be more robust to noisy observations, and that unnecessary changepoints (short lived changes, outliers etc.) are not inferred in this case. For the final example, the methods were applied to some simulated stochastic volatility data. Performance of the approach was promising in this case, however, there was difficulty in detecting some smaller changes.

It is worth noting again that RJMCMC would be practically infeasible for the data models considered here. This is since in addition to the issue of efficient sampling from hierarchical GMRF models (see Rue & Held (2005)), there is also segmentation of the data. Thus adding a new changepoint would require designing a reversible move between proposed and current field hyperparameters and in addition resampling field elements. Making moves of this type which exhibit good mixing would be challenging, and further diagnosing convergence would be difficult with the chains possibly requiring very long run times. This gives the approximate approach even more of an advantage.

This is true especially in the case of models which require good corresponding proposal densities to perform well when it comes to MCMC, such as stochastic volatility models.

Overall, this paper has explored a promising new direction for estimation of changepoint models by creating a hybrid of two popular methods in their respective fields, namely INLAs in the GMRF field of study, and filtering recursions for sequential changepoint model estimation. Other data models are possible which have not been applied to any of the examples in this paper. For example, it is possible to have higher order Markov dependencies for random walk fields in the R-INLA package. Zero inflated Poisson and Binomial data models are also possible.

Acknowledgments

The authors would like to thank the Editor, Associate Editor and two anonymous reviewers for their attentive reading of the manuscript and valuable suggestions which improved the presentation of the ideas in the paper. The first author would like to dedicate his work on this paper to the memory Cian Costello (1984-2010) who is enormously missed as a colleague and friend.

References

Boys, R. J. & Henderson, D. A. (2004), 'A Bayesian Approach to DNA Sequence Segmentation', *Biometrics* **60**, 573–588.

Carlin, B. P., Gelfand, A. E. & Smith, A. F. M. (1992), 'Hierarchical Bayesian Analysis of Changepoint Problems', *Applied Statistics* **2**, 389–405.

Chib, S. (1998), 'Estimation and comparison of multiple change-point models', *Journal of Econometrics* **86**, 221–241.

Cox, D. R. & Lewis, P. A. W. (1966), *The Statistical Analysis of Series of Events*, Methuen, London.

Fearnhead, P. (2005), 'Exact Bayesian Curve Fitting and Signal Segmentation', *IEEE Transactions on Signal Processing* **53**, 2160–2166.

Fearnhead, P. (2006), 'Exact and efficient Bayesian inference for multiple changepoint problems', *Statistics and Computing* **16**, 203–213.

Fearnhead, P. & Clifford, P. (2003), 'On-Line Inference for Hidden Markov Models via Particle Filters', *Journal of the Royal Statistical Society, Series B* **65**, 887–899.

Fearnhead, P. & Liu, Z. (2010), 'Efficient Bayesian analysis of multiple changepoint models with dependence across segments', *Statistics and Computing* . To appear.

Green, P. (1995), 'Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model determination', *Biometrika* **82**, 711–732.

Jarrett, R. G. (1979), 'A note on the intervals between coal-mining disasters', *Biometrika* **66**, 191–193.

Ó Ruanaidh, J. J. K. & Fitzgerald, W. J. (1996), *Numerical Bayesian Methods applied to Signal Processing*, Springer, New York.

Raftery, A. E. & Akman, V. E. (1986), 'Bayesian Analysis of a Poisson Process with a Change-Point', *Biometrika* **73**, 85–89.

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Vol. 104 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)', *Journal of the Royal Statistical Society, Series B* **71**, 319–392.

Scott, S. L. (2002), 'Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century', *Journal of the American Statistical Association* **97**, 337–351.

Wyse, J. & Friel, N. (2010), 'Simulation-based Bayesian analysis for multiple changepoints', *Available at http://arxiv.org/abs/1011.2932* .

Yang, T. Y. & Kuo, L. (2001), 'Bayesian Binary Segmentation Procedure for a Poisson Process with Multiple Changepoints', *Journal of Computational and Graphical Statistics* **10**, 772–785.