NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Approximate simulation free multiple changepoint analysis with Gaussian Markov random field segment models

by

Jason Wyse, Nial Friel and Håvard Rue

PREPRINT
STATISTICS NO. 15/2010

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL
http://www.math.ntnu.no/preprint/statistics/2010/S15-2010.pdf

Håvard Rue has homepage: http://www.math.ntnu.no/∼hrue
E-mail: hrue@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and
Technology, N-7491 Trondheim, Norway.

Approximate simulation free multiple changepoint analysis with Gaussian Markov random field segment models

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

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

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

November 2010

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 model. 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 changepoint 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 there is independent data 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 model 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 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.

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, 5 and 6 analyze real data examples; analysis of data arising from comparative genomic hybridization studies; 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 is analyzed with a model that allows for dependency between adjacent data points, such that the dependency relation may change across segments. Section 7 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. In some experimentation in the first authors PhD thesis, it was 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 a given number of changepoints, 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

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

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 $\Theta = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{k+1}^{\mathrm{T}})^{\mathrm{T}}$ are the segment hyperparameters. Independent priors are taken on the members of Θ and the changepoints. The prior taken on changepoints is assumed to have the product form

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

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

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

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 $L_j^{(k)}(t) = \Pr(\mathbf{y}_{t:n}|\tau_j = t-1 \text{ and } k \text{ changepoints})$. It is possible to compute $L_j^{(k)}(t)$ in a backward recursion;

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

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

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

2.2 Choice of changepoint prior and computational cost

It will be necessary to compute k for a range of values, say $k = 0, \dots, K$ 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 $\{1, \dots, n-1\}$ without replacement. Doing this gives

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

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

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

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;

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

This is used for the forward simulation of changepoints once the backward recursions have been computed. It can also be 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; 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 usually would not make sense to fit a GMRF model to a very small amount of data. 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 use all the data, but to do recursions on a smaller portion of it, in order to approximate the full recursions (3). What is meant by this is that the recursion is not computed at every data point which takes $O(n^2)$ computation. Generally if segments have a reasonable duration, changpoints can be detected in the region where they have occurred. The change in regime will be detectable for a period after the actual changepoint position, possibly until a time is reached where the support for a one segment model may be greater than that for a two segment model. An analysis using RFRs only 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 $i < j$. For convenience, define $t_0 = 0$ and $t_{N+1} = n$. So to clarify, the assumption is that if there is a changepoint between t_i and t_{i+2} it can be detected at t_{i+1}. The spacing of the t_i is clearly 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 for the reduced time index set to cover the entire data, consequently increasing the computation time. The most natural choice is to take equally spaced points if there is little prior knowledge of where changepoints occur. This corresponds to $t_i = ig$ 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 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.

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

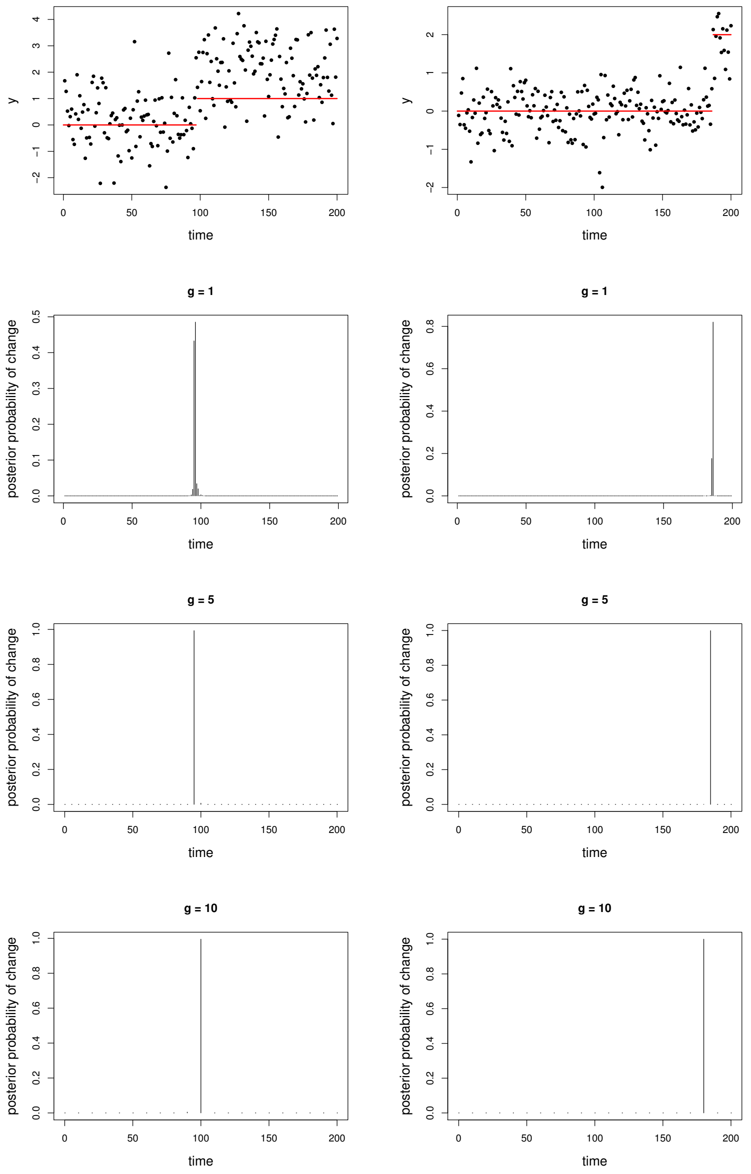
$$R_j^{(k)}(r) = \Pr(\mathbf{y}_{t_r+1:n} | \tau_{j-1} = t_r, k \text{ changepoints}).$$

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 $r = N - k + j - 1, \ldots, j + 1$

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



(a) (b)

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.

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

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

Once the grid spacing g is not too large, 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

Following detection of 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 mean 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.

3.1.4 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 $j \in \{$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 within a segment. 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.

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 2.5GHz 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

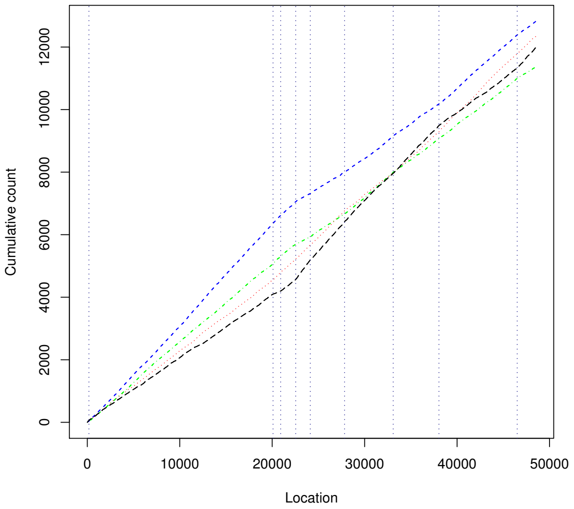


Figure 2: Cumulative counts of A,C,G,T for the DNA data. Identified changepoints are overlain (vertical lines).

g	1	5	10	15	20	25
Time taken (s)	1816.73	76.96	19.22	9.47	5.86	4.03
Changepoints	–	175	175	175	175	175
	20091	20100	20091	20091	20091	20091
	20919	20919	20919	20919	20919	20919
	22546	22584	22545	22545	22545	22545
	24118	24118	24118	24118	24118	24118
	27830	27830	27830	27830	27830	27830
	–	31225	31225	31225	31225	–
	33099	33100	33100	33100	33100	33088
	38036	38035	38048	38010	38035	38035
	46535	46535	46535	46535	–	46500

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.

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 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.

4 CGH studies of Coriel cell lines

Comparative genomic hybridization or CGH studies are used to detect chromosomal aberrations in the genome in tumor tissue. Two tissue samples, one tumor and the other healthy, are dyed with different fluorochromes (red and green). The two samples are then mixed together. Aberrations present in the tumor DNA are detected by examining the colour of the fluorescence emitted by the mixture of the two samples. A yellow fluorescence indicates that there have been no amplifications or deletions in the tumor sample. If the healthy tissue has been dyed red or green however, and the mixture of the samples emits a red or green fluorescence, then there has been chromosomal aberrations in the tumor tissue. After dying and mixing the tissue samples the emitted fluorescence is translated into DNA copy number.

The data studied here is chromosome 11 of Coriel.05296 which has been previously studied by Erdman & Emerson (2007) and Fridlyand, Snijders, Pinkel, Albertson & Jain (2004). The data is available from the R package bcp (Erdman & Emerson 2007). In this type of application, analysis is usually carried out on the log-to-base-two ratio of the red-green intensities obtained from the DNA copy numbers from the fluorescence experiments. The data is shown in Figure 3 for chromosome 11 ($n = 185$). The task is to detect changepoints in this series. There has been some pre-processing of this particular data to remove points with a negligible level of intensity and to correct for background noise. The specifics of this pre-processing are described in Section 3.1 of Erdman & Emerson (2007). Even after pre-processing, this data can still be prone to outliers or "short-lived changes" which can be attributable to false signals or a true signal on a single strand of DNA.

Erdman & Emerson (2007) compare different changepoint analyses for this data. They use models which do not explicitly account for dependence in the chromosome. Fridlyand et al. (2004) use a Hidden Markov Model to account for the spatial dependency in the copy number values. Following Fridlyand et al. (2004), a GMRF model is fitted to account for dependency in the series of copy numbers. The observational data is assumed Gaussian with some variance $\sigma^2_{\mathbf{y}}$ and mean μ given by identity link to an AR(1) field. To be more specific,

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2_{\mathbf{y}})$$

where $\mu_i = \alpha + x_i$. The parameter α is an intercept and

$$x_i = \phi x_{i-1} + \varepsilon_i, \quad i = 2, \ldots, n$$

where $\varepsilon_i \sim_{\text{iid}} \mathrm{N}(0, \sigma_{\mathbf{x}}^2)$. The definition is completed by assuming the marginal distribution of x_1 is $\mathrm{N}(0, \sigma_{\mathbf{x}}^2/(1-\phi^2))$.

There are a few advantages with this model; the mean copy number is modelled at each location by the field; the model expoits the spatial dependence along the chromosome; this model will be more robust to outliers, as the field can model extra intra segment variability. In addition to allowing dependence between neighbouring locations along the chromosome, the approach adopted allows for changes in this dependence pattern across different segments through the persistence parameter (ϕ) of the AR(1) field. A drawback of the approach in general is that both changepoints corresponding to small segments (with length less than about three) may not be detected or correctly located. In many applications however, this is usually not an issue, as segments of duration less than three would generally not be expected. For a discussion on minimum segment duration and priors in this context, see Girón, Moreno & Casella (2007).

Using the R-INLA package requires choosing parameters for the priors of $\sigma_{\mathbf{y}}^2, \sigma_{\mathbf{x}}^2$ and ϕ. A Gamma prior is used for $\sigma_{\mathbf{y}}^{-2}$ and $\sigma_{\mathbf{x}}^{-2}$. The prior on ϕ is specified through $\kappa = \operatorname{logit}\left(\frac{1+\phi}{2}\right)$ where $\kappa \sim \mathrm{N}(\mu_\kappa, \sigma_\kappa^2)$. Priors which are too diffuse could cause problems in approximating marginal likelihoods for small segments, or may demand lots of gridding in $\boldsymbol{\theta}$ to get accurate approximations. This problem is not exclusive to INLAs. As noted by Kass & Raftery (1995) Bayes factors (and marginal likelihoods), tend to be sensitive to the choice of priors on the parameters. Priors were chosen to mimic the behaviour of the data. These were

$$\begin{aligned}
\sigma_{\mathbf{y}}^{-2} &\sim \mathrm{Gamma}(75, 0.5) \\
\sigma_{\mathbf{x}}^{-2} &\sim \mathrm{Gamma}(15, 0.1) \\
\kappa &\sim \mathrm{N}(2, 1).
\end{aligned}$$

The prior on changepoints was taken to be uniform on the integers $\{0, \ldots, 5\}$.

The results were obtained by running R-INLA on a 2.5GHz processor. The value of g was taken to be 5. In performing the calculations in the INLA package, one may use the mode of $\boldsymbol{\theta}$ from the previous iteration as an initial value for the optimization in the next iteration. This should speed up the convergence of the Newton-Raphson scheme for finding the modal configuration of the GMRF. The approximate marginal likelihoods took just over 10 minutes to compute. There was overwhelming support for a two changepoint model. The posterior probability for this was 0.996. Changepoints were found at 50 and 65. A refined search then moved these to 51 and 66. Conditional on these changepoints, inference was performed for the segment fields. This is shown in the bottom of Figure 3. It can be seen that the AR(1) field gives a very good fit to the data. It is possible to assess qualitative differences between the three segments by comparing the approximated marginal posteriors of $\sigma_{\mathbf{y}}^{-2}, \sigma_{\mathbf{x}}^{-2}$ and ϕ (from INLAs) shown in Figure 4. It appears that the segment between 51 and 66 has more posterior support for a larger persistence parameter (mode is about 0.75 compared with 0.375). This segment also appears to have larger variance for the observations and the field, due to the noisier observations. Overall the segment from 66 to 185 is less noisy. This could be due to the larger number of data points used to fit the field.

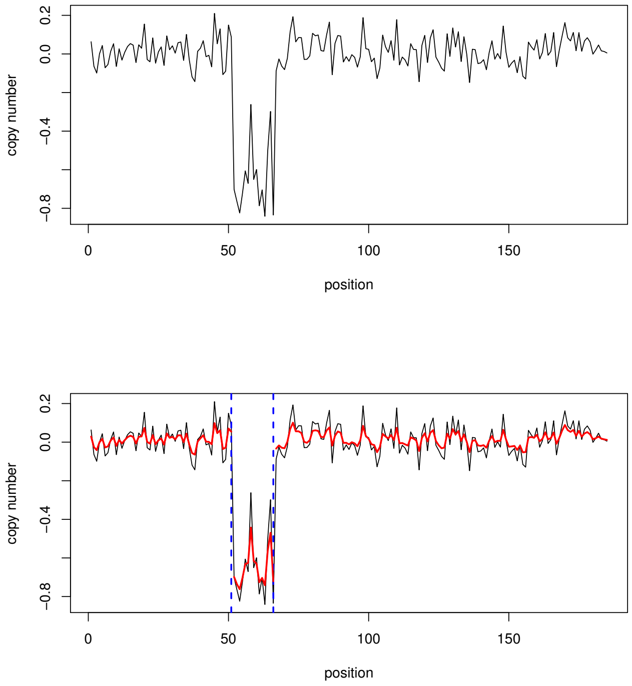


Figure 3: Top: Log to base two of the red-green intensity ratio for Chromosome 11 of Coriell.05296 along the cell line. Note that the horizontal axis here is not scaled identically to that shown in Figure 1 of Erdman and Emerson (2008). (this does not affect the results). Bottom: Inferred changepoints (blue dashed vertical lines) and latent AR(1) field (solid red) using an RFR analysis and INLAs to estimate the field conditional on the detected changepoints.

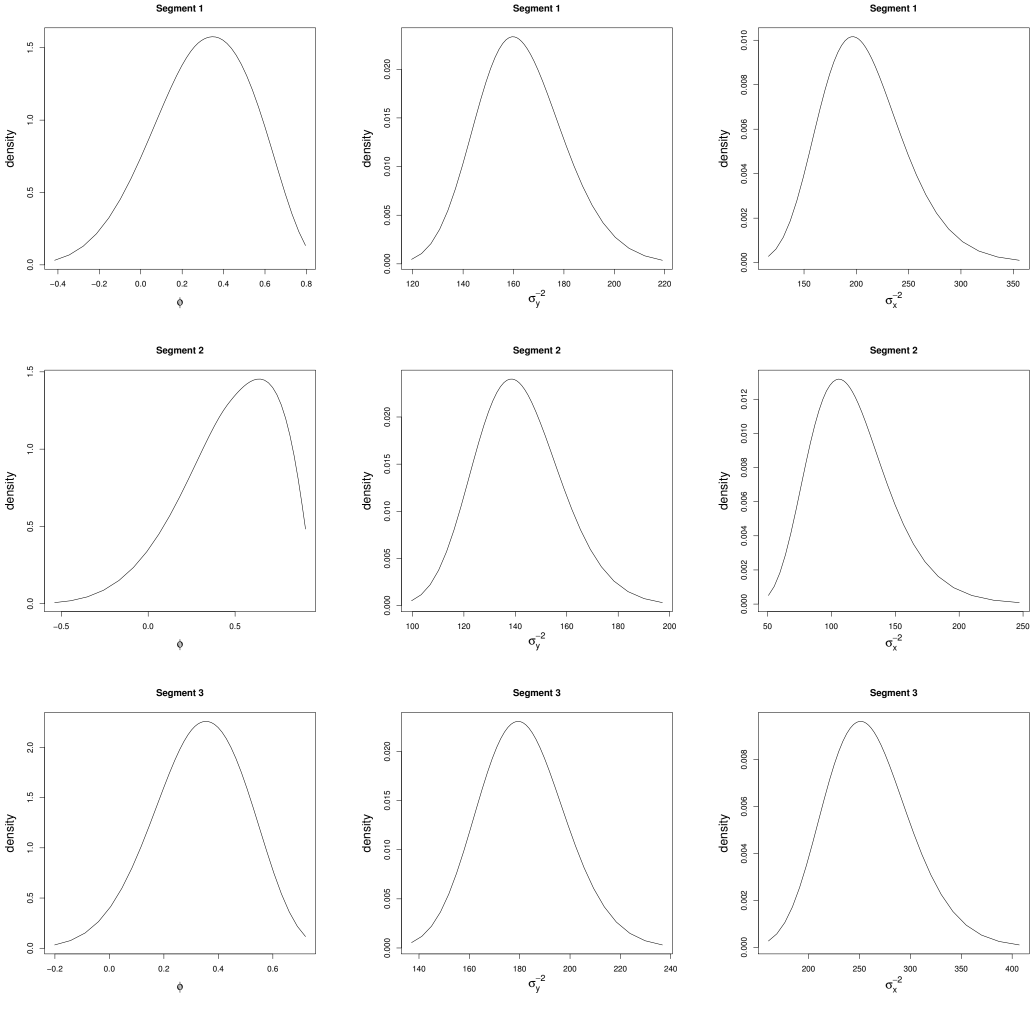


Figure 4: Approximate marginal posterior densities of $\phi, \sigma_{\mathbf{x}}^{-2}$ and $\sigma_{\mathbf{y}}^{-2}$ for each of the three segments.

5 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 in the same way as the Coriel example by choosing them to mimic the behaviour of the data. The priors chosen were

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

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 5 (a) shows the posterior distribution of the number of changepoints for the AR(1) latent field model. A two changepoint

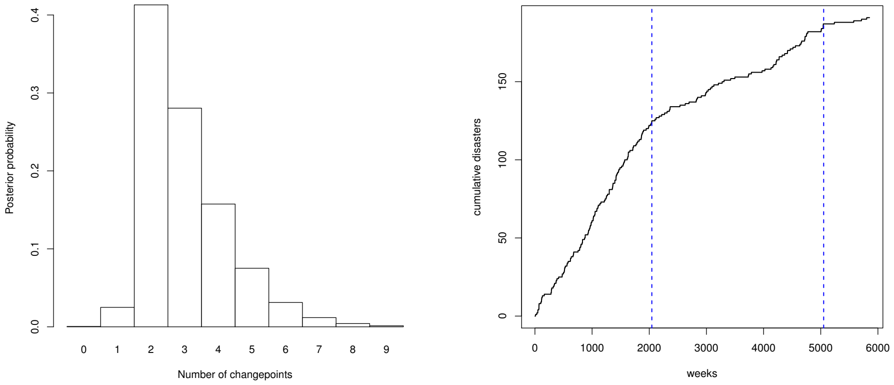


Figure 5: Coal mining data: results from an anlysis using INLAs and $g = 50$. The figure on the left the posterior distribution of the number changes while that on the right shows the cumulative counts of disasters and the changepoints indicated (blue dashed line).

model is most likely, *a posteriori*. Figure 5 (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 6, 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 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 7 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.

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

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

The different models are characterized by model assumptions and consequently the way

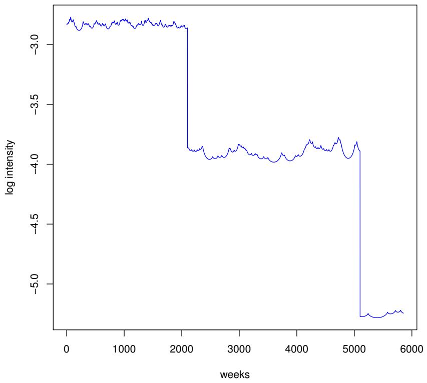


Figure 6: Coal mining data: Inferred log intensity by week.

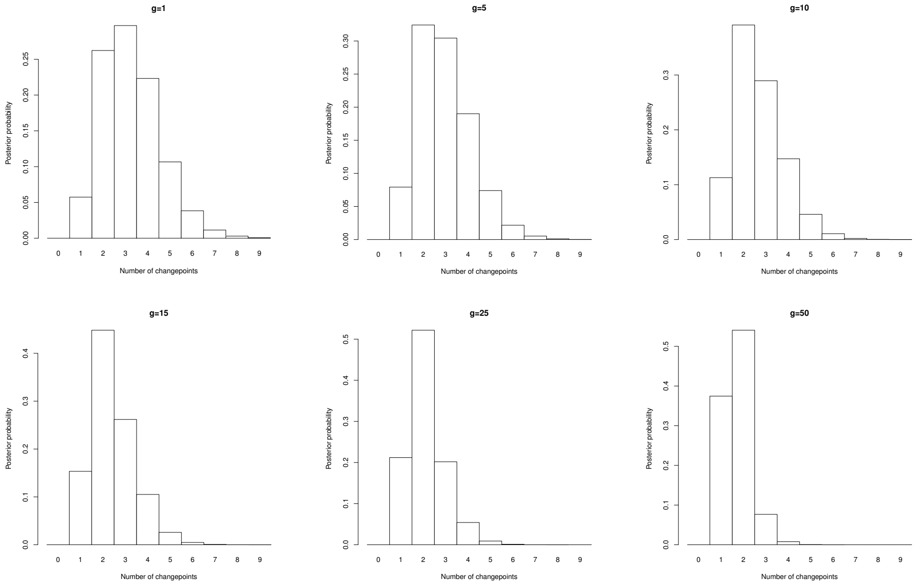


Figure 7: Investigating approximation error in RFRs; results from analyses of coal mining disasters with different values of g using the model from Fearnhead (2006).

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

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

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 6.

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

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

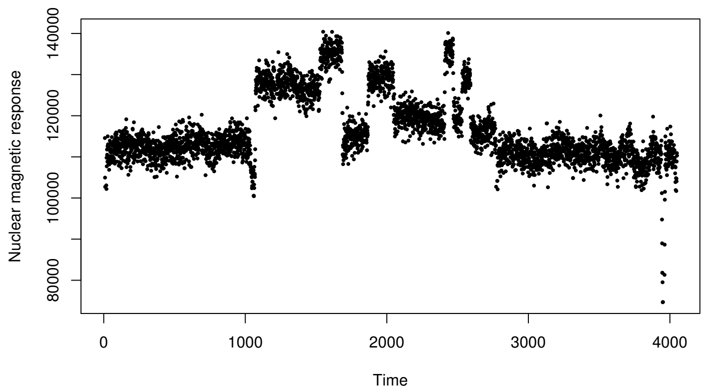


Figure 8: Well-log data. Observations are the nuclear magnetic response recorded by a probe being lowered into a bore-hole in the Earth's surface.

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

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

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.

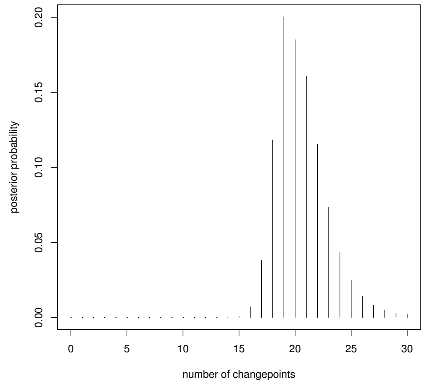


Figure 9: Posterior of the number of changepoints for the Well-log data fitting an AR(1) GMRF model. This suggests the most likely number of changepoints *a posteriori* is 19.

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 9 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 10. 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 10). It can be seen that the independent model is sensitive to changes in the mean and is conservative when 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.

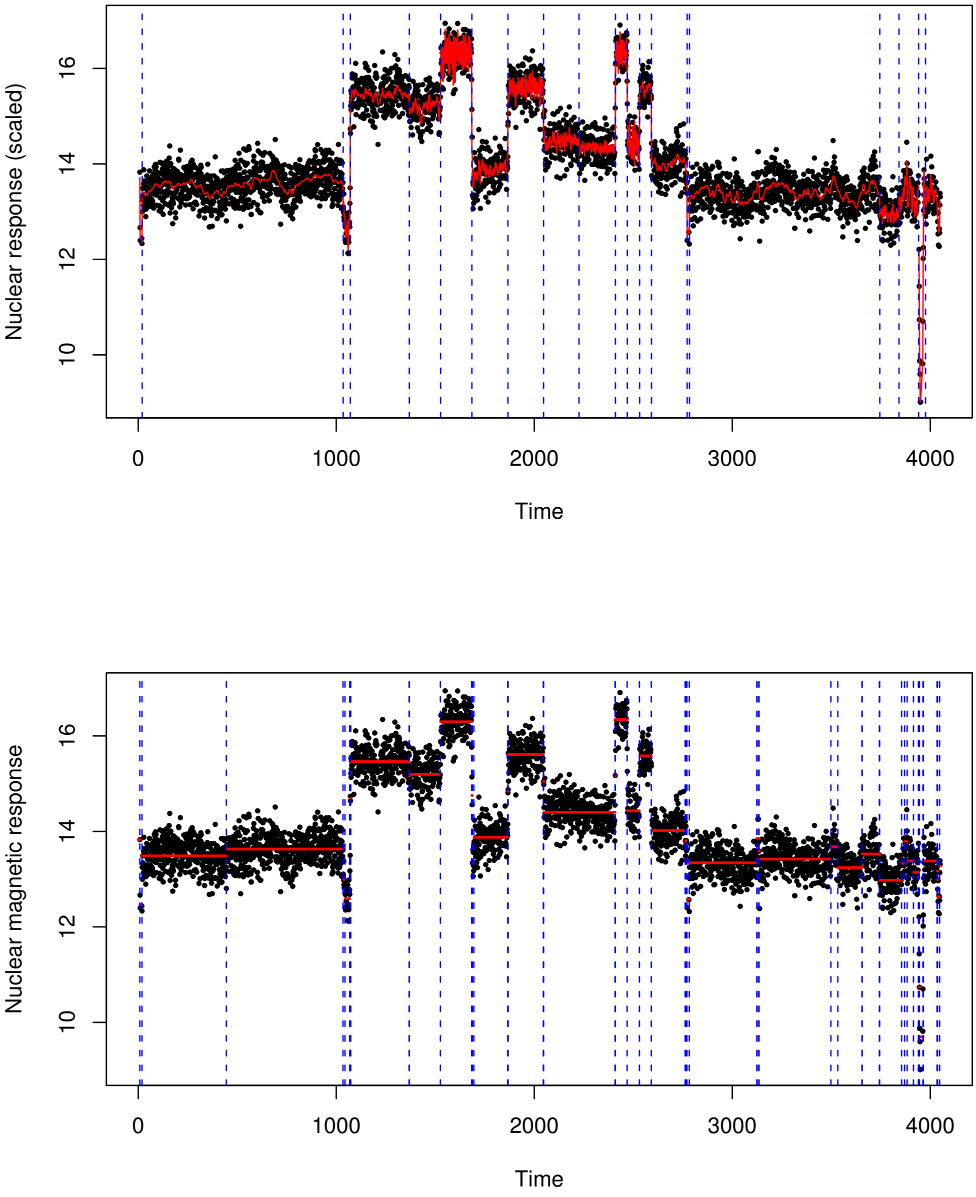


Figure 10: Well-log data: results from RFRs and INLA (top) and independent data model.

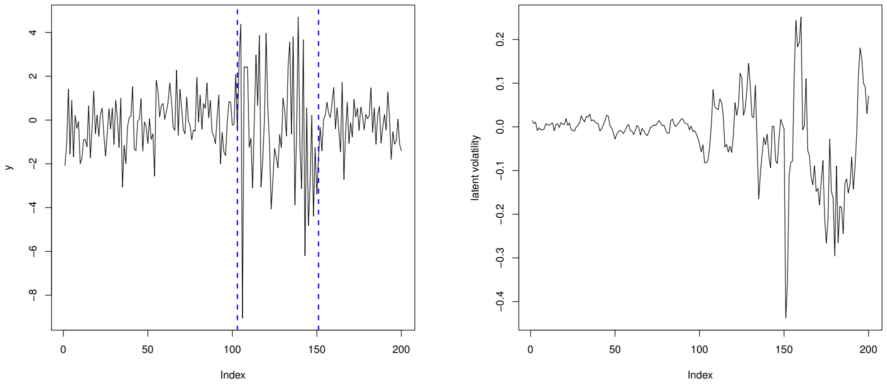


Figure 11: Stochastic volatility data: Simulated observed data with changepoints indicated with blue dashed line (left) and log of simulated latent volatilities (right).

Parameter	Segment 1	Segment 2	Segment 3
Length	100	50	50
$\log \beta$	0	2	0
ϕ	0.8	0.9	0.7
$\sigma^2_{\mathbf{x}}$	0.01^2	0.05^2	0.09^2

Table 2: Stochastic volatility: parameters used to simulate data.

7 Stochastic volatility data

In this example it is demonstrated how INLAs can be used with RFRs to estimate changepoint models where the segment observations are assumed to arise from a stochastic volatility model.

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 $\exp\{\log \beta\}$ 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.

Simulated data was used to test out the methods. This had two changepoints and is described in Table 2. It is shown in Figure 11 along with a plot of the generated latent volatilities.

Prior parameters for computing the segment marginal likelihoods were roughly

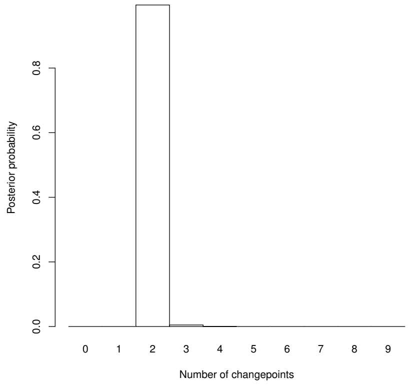


Figure 12: Stochastic volatility data: Posterior distribution of the number of changes.

guessed based by applying INLAs to the entire data. The priors were

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

The required computations took about 5 minutes using a reduced time index set with equal spacing and $g = 5$. A two changepoint model had almost all of the posterior weight, although there was minor support for three changepoints. See Figure 12. The approximation works well in this situation because the change is quite noticeable.

In some other simulated data which the methods were applied to, it was found that the approach was poor at detecting any changepoints when there was only a small (or no) change in the intercept of the latent volatilities $\log \beta$. It was desired that smaller changes, for example, a change in just the persistence parameters across segments, could be detected. This did not seem to be the case however (results not shown).

8 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 four different data examples. The first of these involved detecting changepoints in DNA copy number in CGH studies of Coriel cell lines. The second example 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 third example was an analysis the Well-log data ofÓ 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 was satisfactory when changes were large, but it was noted that more subtle changes in the underlying segments were not detected. This is an area for improvement of the proposed methodology.

It is worth noting again that RJMCMC would be practically infeasible for the data models considered here. 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.

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.

Erdman, C. & Emerson, J. W. (2007), 'bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems', *Journal of Statistical Software* **23**(3).

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.

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D. G. & Jain, A. N. (2004), 'Hidden Markov Models Approach to the Analysis of Array CGH', *Journal of Multivariate Analysis* **90**, 132–153.

Girón, F. J., Moreno, E. & Casella, G. (2007), Objective Bayesian Analysis of Multiple Changepoints for Linear Models, *in* 'Bayesian Statistics 8', Oxford University Press, pp. 227–252.

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.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes Factors', *Journal of the American Statistical Association* **90**, 773–795.

Ó Ruanaidh, J. J. K. & Fitzgerald, W. J. (1996), *Numerical Bayesian Mehtods 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., 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.

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.