Bayesian Online Changepoint Detection

Ryan Prescott Adams
Cavendish Laboratory
Cambridge CB3 0HE
United Kingdom

David J.C. MacKay
Cavendish Laboratory
Cambridge CB3 0HE
United Kingdom

Abstract

Changepoints are abrupt variations in the generative parameters of a data sequence. Online detection of changepoints is useful in modelling and prediction of time series in application areas such as finance, biometrics, and robotics. While frequentist methods have yielded online filtering and prediction techniques, most Bayesian papers have focused on the retrospective segmentation problem. Here we examine the case where the model parameters before and after the changepoint are independent and we derive an online algorithm for exact inference of the most recent changepoint. We compute the probability distribution of the length of the current "run," or time since the last changepoint, using a simple message-passing algorithm. Our implementation is highly modular so that the algorithm may be applied to a variety of types of data. We illustrate this modularity by demonstrating the algorithm on three different real-world data sets.

1 INTRODUCTION

Changepoint detection is the identification of abrupt changes in the generative parameters of sequential data. As an online and offline signal processing tool, it has proven to be useful in applications such as process control [1], EEG analysis [5, 2, 17], DNA segmentation [6], econometrics [7, 18], and disease demographics [9].

Frequentist approaches to changepoint detection, from the pioneering work of Page [22, 23] and Lorden [19] to recent work using support vector machines [10], offer online changepoint detectors. Most Bayesian approaches to changepoint detection, in contrast, have been offline and retrospective [24, 4, 26, 13, 8]. With a few exceptions [16, 20], the Bayesian papers on changepoint detection focus on segmentation and techniques to generate samples from the posterior distribution over changepoint locations.

In this paper, we present a Bayesian changepoint detection algorithm for online inference. Rather than retrospective segmentation, we focus on causal predictive filtering; generating an accurate distribution of the next unseen datum in the sequence, given only data already observed. For many applications in machine intelligence, this is a natural requirement. Robots must navigate based on past sensor data from an environment that may have abruptly changed: a door may be closed now, for example, or the furniture may have been moved. In vision systems, the brightness change when a light switch is flipped or when the sun comes out.

We assume that a sequence of observations $x_1, x_2, \ldots, x_T$ may be divided into non-overlapping *product partitions* [3]. The delineations between partitions are called the changepoints. We further assume that for each partition ρ, the data within it are i.i.d. from some probability distribution $P(x_t \,|\, \boldsymbol{\eta}_\rho)$. The parameters $\boldsymbol{\eta}_\rho$, $\rho = 1, 2, \ldots$ are taken to be i.i.d. as well. We denote the contiguous set of observations between time a and b inclusive as $\boldsymbol{x}_{a:b}$. The discrete *a priori* probability distribution over the interval between changepoints is denoted as $P_{\mathsf{gap}}(g)$.

We are concerned with estimating the posterior distribution over the current "run length," or time since the last changepoint, given the data so far observed. We denote the length of the current run at time t by r_t. We also use the notation $\boldsymbol{x}_t^{(r)}$ to indicate the set of observations associated with the run r_t. As r may be zero, the set $\boldsymbol{x}^{(r)}$ may be empty. We illustrate the relationship between the run length r and some hypothetical univariate data in Figures 1(a) and 1(b).

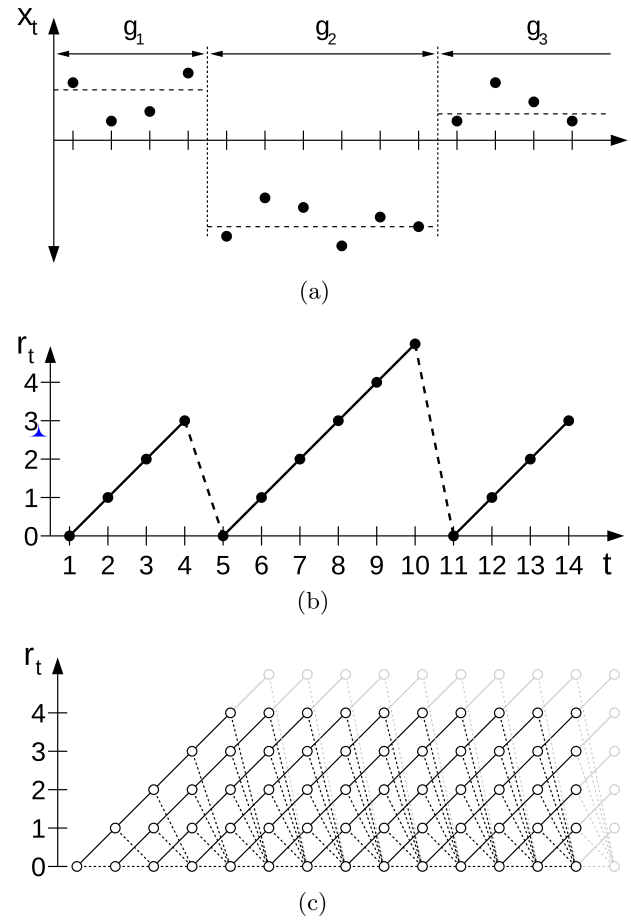

Figure 1: This figure illustrates how we describe a changepoint model expressed in terms of run lengths. Figure 1(a) shows hypothetical univariate data divided by changepoints on the mean into three segments of lengths $g_1 = 4$, $g_2 = 6$, and an undetermined length g_3. Figure 1(b) shows the run length r_t as a function of time. r_t drops to zero when a changepoint occurs. Figure 1(c) shows the trellis on which the message-passing algorithm lives. Solid lines indicate that probability mass is being passed "upwards," causing the run length to grow at the next time step. Dotted lines indicate the possibility that the current run is truncated and the run length drops to zero.

2 RECURSIVE RUN LENGTH ESTIMATION

We assume that we can compute the predictive distribution conditional on a given run length r_t. We then integrate over the posterior distribution on the current run length to find the marginal predictive distribution:

$$P(x_{t+1} \mid \boldsymbol{x}_{1:t}) = \sum_{r_t} P(x_{t+1} \mid r_t, \boldsymbol{x}_t^{(r)}) P(r_t \mid \boldsymbol{x}_{1:t}) \quad (1)$$

To find the posterior distribution

$$P(r_t \mid \boldsymbol{x}_{1:t}) = \frac{P(r_t, \boldsymbol{x}_{1:t})}{P(\boldsymbol{x}_{1:t})}, \quad (2)$$

we write the joint distribution over run length and observed data recursively.

$$\begin{aligned} P(r_t, \boldsymbol{x}_{1:t}) &= \sum_{r_{t-1}} P(r_t, r_{t-1}, \boldsymbol{x}_{1:t}) \\ &= \sum_{r_{t-1}} P(r_t, x_t \mid r_{t-1}, \boldsymbol{x}_{1:t-1}) P(r_{t-1}, \boldsymbol{x}_{1:t-1}) \quad (3) \\ &= \sum_{r_{t-1}} P(r_t \mid r_{t-1}) P(x_t \mid r_{t-1}, \boldsymbol{x}_t^{(r)}) P(r_{t-1}, \boldsymbol{x}_{1:t-1}) \end{aligned}$$

How is this derived? Shouldn't this term be 1?

Note that the predictive distribution $P(x_t \mid r_{t-1}, \boldsymbol{x}_{1:t})$ depends only on the recent data $\boldsymbol{x}_t^{(r)}$. We can thus generate a recursive message-passing algorithm for the joint distribution over the current run length and the data, based on two calculations: 1) the prior over r_t given r_{t-1}, and 2) the predictive distribution over the newly-observed datum, given the data since the last changepoint.

2.1 THE CHANGEPOINT PRIOR

The conditional prior on the changepoint $P(r_t \mid r_{t-1})$ gives this algorithm its computational efficiency, as it has nonzero mass at only two outcomes: the run length either continues to grow and $r_t = r_{t-1}+1$ or a changepoint occurs and $r_t = 0$.

$$P(r_t \mid r_{t-1}) = \begin{cases} H(r_{t-1}+1) & \text{if } r_t = 0 \\ 1 - H(r_{t-1}+1) & \text{if } r_t = r_{t-1} + 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The function $H(\tau)$ is the *hazard function*. [11].

$$H(\tau) = \frac{P_{\mathsf{gap}}(g{=}\tau)}{\sum_{t=\tau}^{\infty} P_{\mathsf{gap}}(g{=}t)} \quad (5)$$

What exactly Hazard function is

In the special case is where $P_{\mathsf{gap}}(g)$ is a discrete exponential (geometric) distribution with timescale λ, the process is memoryless and the hazard function is constant at $H(\tau) = 1/\lambda$.

Figure 1(c) illustrates the resulting message-passing algorithm. In this diagram, the circles represent run-length hypotheses. The lines between the circles show recursive transfer of mass between time steps. Solid lines indicate that probability mass is being passed "upwards," causing the run length to grow at the next time step. Dotted lines indicate that the current run is truncated and the run length drops to zero.

2.2 BOUNDARY CONDITIONS

A recursive algorithm must not only define the recurrence relation, but also the initialization conditions. We consider two cases: 1) a changepoint occurred *a priori* before the first datum, such as when observing a game. In such cases we place all of the probability mass for the initial run length at zero, i.e. $P(r_0{=}0) = 1$. 2) We observe some recent subset of the data, such as when modelling climate change. In this case the prior over the initial run length is the normalized *survival function* [11]

Significance of Survival function?

$$P(r_0{=}\tau) = \frac{1}{Z} S(\tau), \tag{6}$$

where Z is an appropriate normalizing constant, and

$$S(\tau) = \sum_{t=\tau+1}^{\infty} P_{\mathsf{gap}}(g{=}t). \tag{7}$$

2.3 CONJUGATE-EXPONENTIAL MODELS

Conjugate-exponential models are particularly convenient for integrating with the changepoint detection scheme described here. Exponential family likelihoods allow inference with a finite number of sufficient statistics which can be calculated incrementally as data arrives. Exponential family likelihoods have the form

$$P(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \boldsymbol{U}(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \tag{8}$$

where

$$A(\boldsymbol{\eta}) = \log \int d\boldsymbol{\eta}\, h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \boldsymbol{U}(\boldsymbol{x})\right). \tag{9}$$

The strength of the conjugate-exponential representation is that both the prior and posterior take the form of an exponential-family distribution over $\boldsymbol{\eta}$ that can be summarized by succinct hyperparameters ν and $\boldsymbol{\chi}$.

$$P(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = \tilde{h}(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \boldsymbol{\chi} - \nu A(\boldsymbol{\eta}) - \tilde{A}(\boldsymbol{\chi}, \nu)\right) \tag{10}$$

We wish to infer the parameter vector $\boldsymbol{\eta}$ associated with the data from a current run length r_t. We denote this run-specific model parameter as $\boldsymbol{\eta}_t^{(r)}$. After finding the posterior distribution $P(\boldsymbol{\eta}_t^{(r)} \mid r_t, \boldsymbol{x}_t^{(r)})$, we can marginalize out the parameters to find the predictive distribution, conditional on the length of the current run.

$$P(x_{t+1} \mid r_t) = \int d\boldsymbol{\eta}\, P(x_{t+1} \mid \boldsymbol{\eta}) P(\boldsymbol{\eta}_t^{(r)}{=}\boldsymbol{\eta} \mid r_t, \boldsymbol{x}_t^{(r)}) \tag{11}$$

1. **Initialize**
$$P(r_0) = \tilde{S}(r) \text{ or } P(r_0{=}0) = 1$$
$$\nu_1^{(0)} = \nu_{\mathsf{prior}}$$
$$\boldsymbol{\chi}_1^{(0)} = \boldsymbol{\chi}_{\mathsf{prior}}$$
2. **Observe New Datum** x_t
3. **Evaluate Predictive Probability**
$$\pi_t^{(r)} = P(x_t \mid \nu_t^{(r)}, \boldsymbol{\chi}_t^{(r)})$$
4. **Calculate Growth Probabilities**
$$P(r_t{=}r_{t-1}{+}1, \boldsymbol{x}_{1:t}) = P(r_{t-1}, \boldsymbol{x}_{1:t-1}) \pi_t^{(r)} (1{-}H(r_{t-1}))$$
5. **Calculate Changepoint Probabilities**
$$P(r_t{=}0, \boldsymbol{x}_{1:t}) = \sum_{r_{t-1}} P(r_{t-1}, \boldsymbol{x}_{1:t-1}) \pi_t^{(r)} H(r_{t-1})$$
6. **Calculate Evidence**
$$P(\boldsymbol{x}_{1:t}) = \sum_{r_t} P(r_t, \boldsymbol{x}_{1:t})$$
7. **Determine Run Length Distribution**
$$P(r_t \mid \boldsymbol{x}_{1:t}) = P(r_t, \boldsymbol{x}_{1:t}) / P(\boldsymbol{x}_{1:t})$$
8. **Update Sufficient Statistics**
$$\nu_{t+1}^{(0)} = \nu_{\mathsf{prior}}$$
$$\boldsymbol{\chi}_{t+1}^{(0)} = \boldsymbol{\chi}_{\mathsf{prior}}$$
$$\nu_{t+1}^{(r+1)} = \nu_t^{(r)} + 1$$
$$\boldsymbol{\chi}_{t+1}^{(r+1)} = \boldsymbol{\chi}_t^{(r)} + \boldsymbol{u}(x_t)$$
9. **Perform Prediction**
$$P(x_{t+1} \mid \boldsymbol{x}_{1:t}) = \sum_{r_t} P(\boldsymbol{x}_{t+1} | \boldsymbol{x}_t^{(r)}, r_t) P(r_t | \boldsymbol{x}_{1:t})$$
10. **Return to Step 2**

Algorithm 1: The online changepoint algorithm with prediction. An additional optimization not shown is to truncate the per-timestep vectors when the tail of $P(r_t|\boldsymbol{x}_{1:t})$ has mass beneath a threshold.

This marginal predictive distribution, while generally not itself an exponential-family distribution, is usually a simple function of the sufficient statistics. When exact distributions are not available, compact approximations such as that described by Snelson and Ghahramani [25] may be useful. We will only address the exact case in this paper, where the predictive distribution associated with a particular current run length is parameterized by $\nu_t^{(r)}$ and $\boldsymbol{\chi}_t^{(r)}$.

$$\nu_t^{(r)} = \nu_{\mathsf{prior}} + r_t \tag{12}$$

$$\boldsymbol{\chi}_t^{(r)} = \boldsymbol{\chi}_{\mathsf{prior}} + \sum_{t' \in r_t} \boldsymbol{u}(x_{t'}) \tag{13}$$

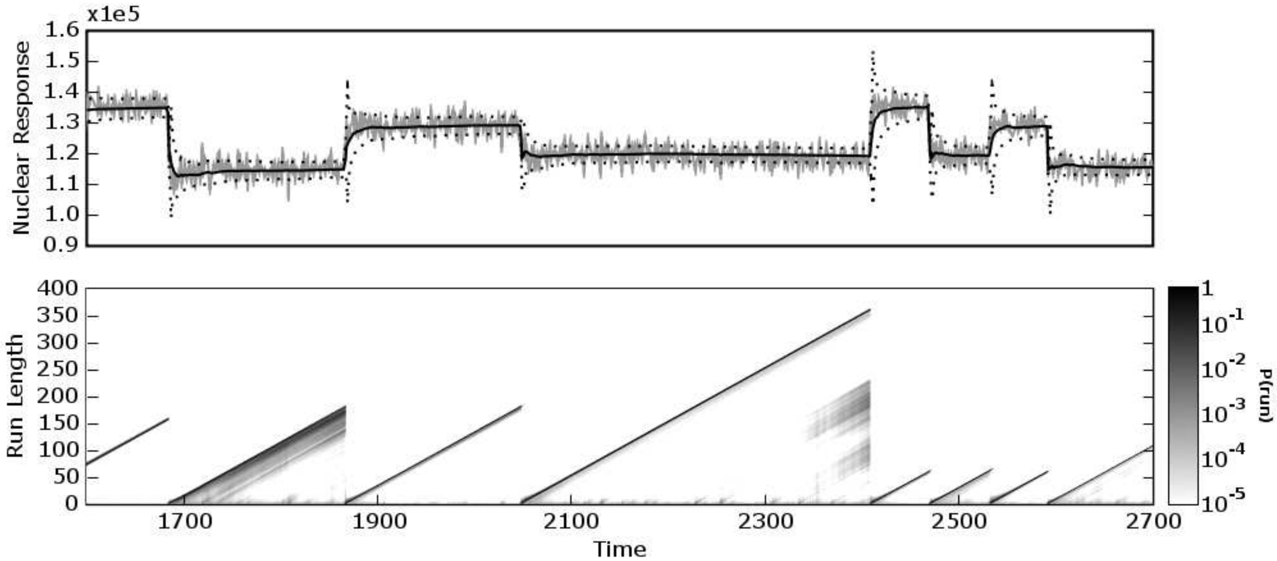

Figure 2: The top plot is a 1100-datum subset of nuclear magnetic response during the drilling of a well. The data are plotted in light gray, with the predictive mean (solid dark line) and predictive 1-σ error bars (dotted lines) overlaid. The bottom plot shows the posterior probability of the current run $P(r_t \mid \boldsymbol{x}_{1:t})$ at each time step, using a logarithmic color scale. Darker pixels indicate higher probability.

2.4 COMPUTATIONAL COST

The complete algorithm, assuming exponential-family likelihoods, is shown in Algorithm 1. The space- and time-complexity per time-step are linear in the number of data points so far observed. A trivial modification of the algorithm is to discard the run length probability estimates in the tail of the distribution which have a total mass less than some threshold, say 10^{-4}. This yields a constant average complexity per iteration on the order of the expected run length $E[r]$, although the worst-case complexity is still linear in the data.

3 EXPERIMENTAL RESULTS

In this section we demonstrate several implementations of the changepoint algorithm developed in this paper. We examine three real-world example datasets. The first case is a varying Gaussian mean from well-log data. In the second example we consider abrupt changes of variance in daily returns of the Dow Jones Industrial Average. The final data are the intervals between coal mining disasters, which we model as a Poisson process. In each of the three examples, we use a discrete exponential prior over the interval between changepoints.

3.1 WELL-LOG DATA

These data are 4050 measurements of nuclear magnetic response taken during the drilling of a well. The data are used to interpret the geophysical structure of the rock surrounding the well. The variations in mean reflect the stratification of the earth's crust. These data have been studied in the context of changepoint detection by Ó Ruanaidh and Fitzgerald [21], and by Fearnhead and Clifford [12].

The changepoint detection algorithm was run on these data using a univariate Gaussian model with prior parameters $\mu = 1.15 \times 10^5$ and $\sigma = 1 \times 10^4$. The rate of the discrete exponential prior, λ_{gap}, was 250. A subset of the data is shown in Figure 2, with the predictive mean and standard deviation overlaid on the top plot. The bottom plot shows the log probability over the current run length at each time step. Notice that the drops to zero run-length correspond well with the abrupt changes in the mean of the data. Immediately after a changepoint, the predictive variance increases, as would be expected for a sudden reduction in data.

3.2 1972-75 DOW JONES RETURNS

During the three year period from the middle of 1972 to the middle of 1975, several major events occurred that had potential macroeconomic effects. Significant among these are the Watergate affair and the OPEC oil embargo. We applied the changepoint detection algorithm described here to daily returns of the Dow Jones Industrial Average from July 3, 1972 to June 30, 1975. We modelled the returns

$$R_t = \frac{p_t^{\mathsf{close}}}{p_{t-1}^{\mathsf{close}}} - 1, \tag{14}$$

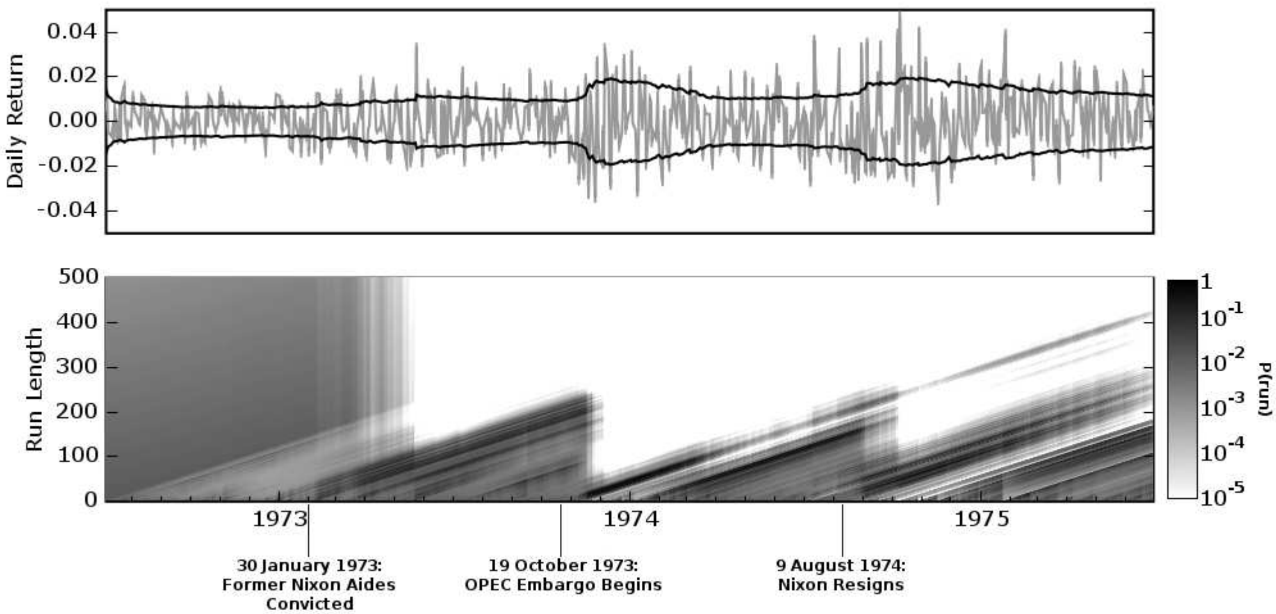

Figure 3: The top plot shows daily returns on the Dow Jones Industrial Average, with an overlaid plot of the predictive volatility. The bottom plot shows the posterior probability of the current run length $P(r_t \mid \boldsymbol{x}_{1:t})$ at each time step, using a logarithmic color scale. Darker pixels indicate higher probability. The time axis is in business days, as this is market data. Three events are marked: the conviction of G. Gordon Liddy and James W. McCord, Jr. on January 30, 1973; the beginning of the OPEC embargo against the United States on October 19, 1973; and the resignation of President Nixon on August 9, 1974.

(where p^{close} is the daily closing price) with a zero-mean Gaussian distribution and piecewise-constant variance. Hsu [14] performed a similar analysis on a subset of these data, using frequentist techniques and weekly returns.

We used a gamma prior on the inverse variance, with $a = 1$ and $b = 10^{-4}$. The exponential prior on changepoint interval had rate $\lambda_{\mathsf{gap}} = 250$. In Figure 3, the top plot shows the daily returns with the predictive standard deviation overlaid. The bottom plot shows the posterior probability of the current run length, $P(r_t \mid \boldsymbol{x}_{1:t})$. Three events are marked on the plot: the conviction of Nixon re-election officials G. Gordon Liddy and James W. McCord, Jr., the beginning of the oil embargo against the United States by the Organization of Petroleum Exporting Countries (OPEC), and the resignation of President Nixon.

3.3 COAL MINE DISASTER DATA

These data from Jarrett [15] are dates of coal mining explosions that killed ten or more men between March 15, 1851 and March 22, 1962. We modelled the data as an Poisson process by weeks, with a gamma prior on the rate with $a = b = 1$. The rate of the exponential prior on the changepoint inverval was $\lambda_{\mathsf{gap}} = 1000$. The data are shown in Figure 4. The top plot shows the cumulative number of accidents. The rate of the Possion process determines the local average of the slope. The posterior probability of the current run length is shown in the bottom plot. The introduction of the Coal Mines Regulations Act in 1887 (corresponding to weeks 1868 to 1920) is also marked.

4 DISCUSSION

This paper contributes a predictive, online interpetation of Bayesian changepoint detection and provides a simple and exact method for calculating the posterior probability of the current run length. We have demonstrated this algorithm on three real-world data sets with different modelling requirements.

Additionally, this framework provides convenient delineation between the implementation of the changepoint algorithm and the implementation of the model. This modularity allows changepoint-detection code to use an object-oriented, "pluggable" type architecture.

Acknowledgements

The authors would like to thank Phil Cowans and Marian Frazier for valuable discussions. This work was funded by the Gates Cambridge Trust.

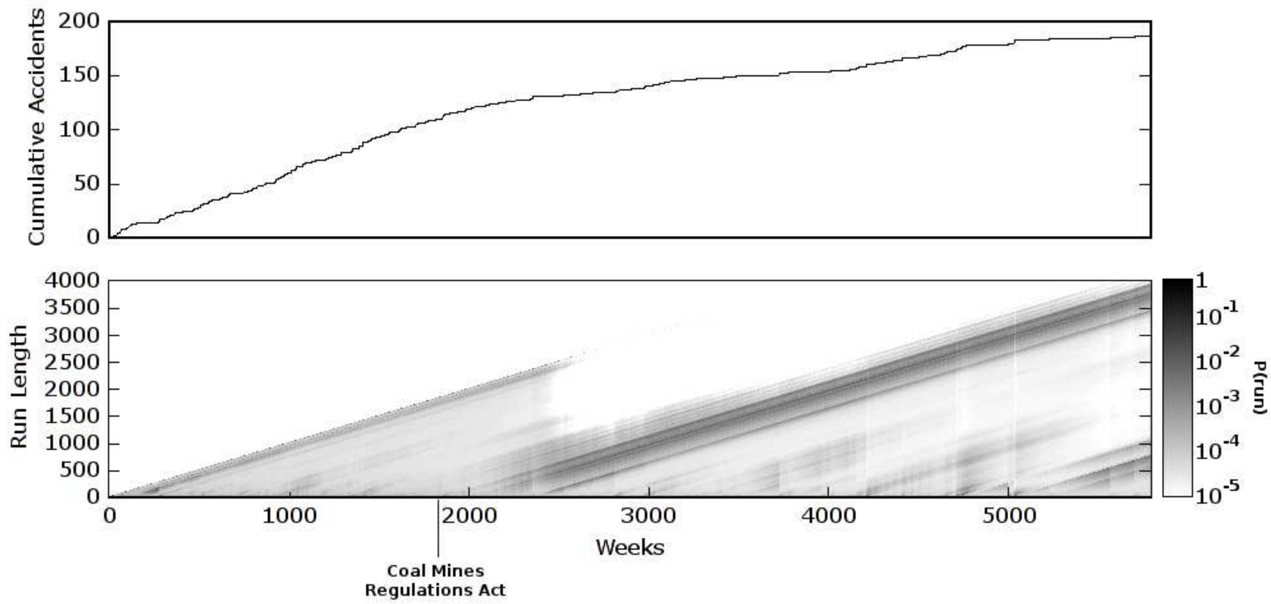

Figure 4: These data are the weekly occurrence of coal mine disasters that killed ten or more people between 1851 and 1962. The top plot is the cumulative number of accidents. The accident rate determines the local average slope of the plot. The introduction of the Coal Mines Regulations Act in 1887 is marked. The year 1887 corresponds to weeks 1868 to 1920 on this plot. The bottom plot shows the posterior probability of the current run length at each time step, $P(r_t \,|\, \boldsymbol{x}_{1:t})$.

References

[1] Leo A. Aroian and Howard Levene. The effectiveness of quality control charts. *Journal of the American Statistical Association*, 45(252):520–529, 1950.

[2] J. S. Barlow, O. D. Creutzfeldt, D. Michael, J. Houchin, and H. Epelbaum. Automatic adaptive segmentation of clinical EEGs,. *Electroencephalography and Clinical Neurophysiology*, 51(5):512–525, May 1981.

[3] D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, 20:260–279, 1992.

[4] D. Barry and J. A. Hartigan. A Bayesian analysis of change point problems. *Journal of the American Statistical Association*, 88:309–319, 1993.

[5] G. Bodenstein and H. M. Praetorius. Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE*, 65(5):642–652, 1977.

[6] J. V. Braun, R. K. Braun, and H. G. Müller. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2):301–314, June 2000.

[7] Jie Chen and A. K. Gupta. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92(438):739–747, June 1997.

[8] Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, October 1998.

[9] D. Denison and C. Holmes. Bayesian partitioning for estimating disease risk, 1999.

[10] F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, August 2005.

[11] Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. Wiley-Interscience, June 2000.

[12] Paul Fearnhead and Peter Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society B*, 65(4):887–899, 2003.

[13] P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, 1995.

[14] D. A. Hsu. Tests for variance shift at an unknown time point. *Applied Statistics*, 26(3):279–284, 1977.

[15] R. G. Jarrett. A note on the intervals between coal-mining disasters. *Biometrika*, 66(1):191–193, 1979.

[16] Timothy T. Jervis and Stuart I. Jardine. Alarm system for wellbore site. United States Patent 5,952,569, October 1997.

[17] A. Y. Kaplan and S. L. Shishkin. Application of the change-point analysis to the investigation of the brain's electrical activity. In B. E. Brodsky and B. S. Darkhovsky, editors, *Non-Parametric Statistical Diagnosis : Problems and Methods*, pages 333–388. Springer, 2000.

[18] Gary M. Koop and Simon M. Potter. Forecasting and estimating multiple change-point models with an unknown number of change points. Technical report, Federal Reserve Bank of New York, December 2004.

[19] G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, December 1971.

[20] J. J. Ó Ruanaidh, W. J. Fitzgerald, and K. J. Pope. Recursive Bayesian location of a discontinuity in time series. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume iv, pages IV/513–IV/516 vol.4, 1994.

[21] Joseph J. K. Ó Ruanaidh and William J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing (Statistics and Computing)*. Springer, February 1996.

[22] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, June 1954.

[23] E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[24] A. F. M. Smith. A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416, 1975.

[25] Edward Snelson and Zoubin Ghahramani. Compact approximations to Bayesian predictive distributions. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 840–847, New York, NY, USA, 2005. ACM Press.

[26] D. A. Stephens. Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, 43:159–178, 1994.