Probabilistic Methods for Time-Series Analysis

Contents

Chapter 1

Analysis of Changepoint Models

Idris A. Eckley, Paul Fearnhead and Rebecca Killick[1]

1.1 Introduction

Many time series are characterised by abrupt changes in structure, such as sudden jumps in level or volatility. We consider change points to be those time points which divide a data set into distinct homogeneous segments. In practice the number of change points will not be known.

The ability to detect changepoints is important for both methodological and practical reasons including: the validation of an untested scientific hypothesis [27]; monitoring and assessment of safety critical processes [14]; and the validation of modelling assumptions [21].

The development of inference methods for change point problems is by no means a recent phenomenon, with early works including [39], [45] and [28]. Increasingly the ability to detect change points quickly and accurately is of interest to a wide range of disciplines. Recent examples of application areas include numerous bioinformatic applications [37, 15] the detection of malware within software [51], network traffic analysis [35], finance [46], climatology [32] and oceanography [34].

In this chapter we describe and compare a number of different approaches for estimating changepoints. For a more general overview of changepoint methods, we refer interested readers to [8] and [11].

The structure of this chapter is as follows. First we introduce the model we focus on. We then describe methods for detecting a single changepoint and methods for detecting multiple changepoints, which will cover both frequentist and Bayesian approaches. For multiple changepoint models the computational challenge of performing inference is to deal with the large space of possible sets of changepoint positions. We describe algorithms that, for the class of models we consider, can perform inference exactly even for large data sets. In Section 1.4 we look at practical issues of implementing these methods, and compare the different approaches, through a detailed simulation study. Our study is based around the problem of detecting changes in the covariance structure of a time-series, and results suggest that Bayesian methods are more suitable for detection of changepoints, particularly for multiple changepoint applications. The study also demonstrates the advantage of using exact inference methods. We end with a discussion.

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF

1.1.1 Model and Notation

Within this chapter we consider the following changepoint models. Let us assume we have time-series data, $y_{1:n} = (y_1, \ldots, y_n)$. For simplicity we assume the observation at each time t, y_t, is univariate – though extensions to multivariate data are straightforward. Our model will have a number of changepoints, m, together with their positions, $\tau_{1:m} = (\tau_1, \ldots, \tau_m)$. Each changepoint position is an integer between 1 and $n-1$ inclusive. We define $\tau_0 = 0$ and $\tau_{m+1} = n$, and assume that the changepoints are ordered so that $\tau_i < \tau_j$ if and only if $i < j$.

The m changepoints will split the data into $m+1$ segments. The ith segment will consist of data $y_{\tau_{i-1}+1:\tau_i}$. For each segment there will be a set of parameters; the parameters associated with the ith segment will be denoted θ_i. We will write the likelihood function as

$$L(m, \tau_{1:m}, \theta_{1:m+1}) = p(y_{1:n}|m, \tau_{1:m}, \theta_{1:m+1}).$$

Here and throughout we use $p(\cdot|\cdot)$ to denote a general conditional density function. Finally we assume conditional independence of data across segments, so that

$$p(y_{1:n}|m, \tau_{1:m}, \theta_{1:m+1}) = \prod_{i=1}^{m+1} p(y_{(\tau_{i-1}+1):\tau_i}|\theta_i).$$

For any segment we will assume we can calculate, either analytically or numerically, the maximum likelihood estimator for the segment parameter. We will denote this by $\hat{\theta}$ or $\hat{\theta}_i$ depending on the context. Thus we have

$$\max_{\theta} p(y_{(\tau_{i-1}+1):\tau_i}|\theta) = p(y_{(\tau_{i-1}+1):\tau_i}|\hat{\theta}).$$

When considering this problem within a Bayesian framework, we will need to introduce priors on both the number and position of changepoints, and on the parameters for each segment. Choice for the former will be discussed below. For the latter, we will assume an exchangeable prior structure. Thus we introduce a family of distributions $p(\theta|\psi)$, parametrised by hyperparameters ψ. Then, conditional on ψ we have $p(\theta_{1:m+1}|\psi) = \prod_{i=1}^{m+1} p(\theta_i|\psi)$. Either we specify ψ, or the model is then completed through an appropriate hyperprior on ψ. Note that the prior, $p(\theta|\psi)$, can be interpreted as describing the variability of the parameters across segments.

For fixed ψ, if we have a segment consisting of observations $y_{s:t}$ for $s < t$, then the segment marginal likelihood is defined as

$$Q(s, t; \psi) = \int p(y_{s:t}|\theta)p(\theta|\psi)\mathrm{d}\theta. \tag{1.1}$$

For the algorithms for Bayesian inference that we focus on, it is important that the marginal likelihoods, $Q(s, t; \psi)$, can be calculated for all s, t and ψ. For many models, this can be done analytically; whilst for others it may be possible to calculate the marginal likelihoods numerically. In most cases, the assumption that we can calculate $Q(s, t; \psi)$ is equivalent to the assumption we can calculate the posterior distribution of the parameter associated with the segment, given the start and end of the segment. Thus in this case, if we can calculate the posterior for the position and number of the changepoints, then we can easily extend this to include the segment parameters as well.

To make this model concrete, we now give an important example which will be used in the simulation studies below.

1.1.2 Example: Piecewise Linear Regression

Assume that for each time-point t we have a d-dimensional covariate $z_t = (z_t^{(1)}, \ldots, z_t^{(d)})$. Our model fits a different linear regression model within each segment. The parameter for each segment consists of the parameters of the linear regressor and the variance of the observations. We denote $\theta_i = (\beta_i^{(1)}, \ldots, \beta_i^{(d)}, \sigma_i^2)$. For segment i, we have $p(y_{(\tau_{i-1}+1):\tau_i}|\theta_i) = \prod_{t=\tau_{i-1}+1}^{\tau_i} p(y_t|\theta_i)$, where, for $t = \tau_{i-1}+1, \ldots, \tau_i$, $Y_t|\theta_i \sim \mathcal{N}\left(\sum_{j=1}^{d} z_t^{(j)}\beta_i^{(j)}, \sigma_i^2\right)$, and $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian random variable, with mean μ and variance σ^2.

Figure 1.1 gives example realisations from these models. Note that special cases of this model include piecewise polynomial models, where $z_t^{(j)} = t^{j-1}$; and, when $d = 0$, changepoint models for the variance of the time-series. Also by letting $z_t^{(j)} = y_{t-j}$ we obtain piecewise auto-regression models. See [41, 12] for more details of these models, and their applications.

Conditional on knowing the segments, inference via maximum likelihood estimation can be performed analytically.

For a Bayesian analysis, we require a prior for θ_i. There are computational advantages in choosing the conjugate prior for this model. If we introduce hyperparameters $\psi = \{a, b, \eta, H\}$, where a and b are scalars, η is a d-dimensional vector, and H is a $d \times d$ matrix, then the conjugate prior is

$$\sigma_i^2|a, b \sim \mathcal{IG}(a, b), \tag{1.2}$$

$$(\beta_i^{(1)}, \ldots, \beta_i^{(d)})|\sigma^2, \eta, H \sim \mathcal{N}(\eta, \sigma^2 H). \tag{1.3}$$

Here $\mathcal{IG}$ denotes an inverse-gamma random variable, and $\mathcal{N}$ a multi-variate normal random variable. Choice of these conjugate priors means that conditional on τ_{i-1} and τ_i, the posterior for θ_i can be calculated analytically – it is from the same inverse-gamma, multi-variate normal family. Also the marginal likelihood for a segment (1.1) can also be calculated analytically [41].

1.2 Single Changepoint Models

We now describe a range of methods for detecting a single changepoint. In each case we will focus on the model introduced above, and just briefly comment on extensions to other models.

1.2.1 Likelihood-ratio based approach

A natural approach to detecting a single changepoint is to view it as performing a hypothesis test. We define the null (H_0) and alternative (H_1) hypotheses for a change as

$$\begin{aligned} H_0 &: \text{No changepoint}, m = 0. \\ H_1 &: \text{A single changepoint}, m = 1. \end{aligned}$$

We now introduce the general likelihood-ratio based approach to test this hypothesis. The potential for using a likelihood based approach to detect changepoints was first proposed by [28] who derives the asymptotic distribution of the likelihood ratio test statistic for a change in the mean within a sequence of normally distributed observations. The likelihood based approach has also been extended to

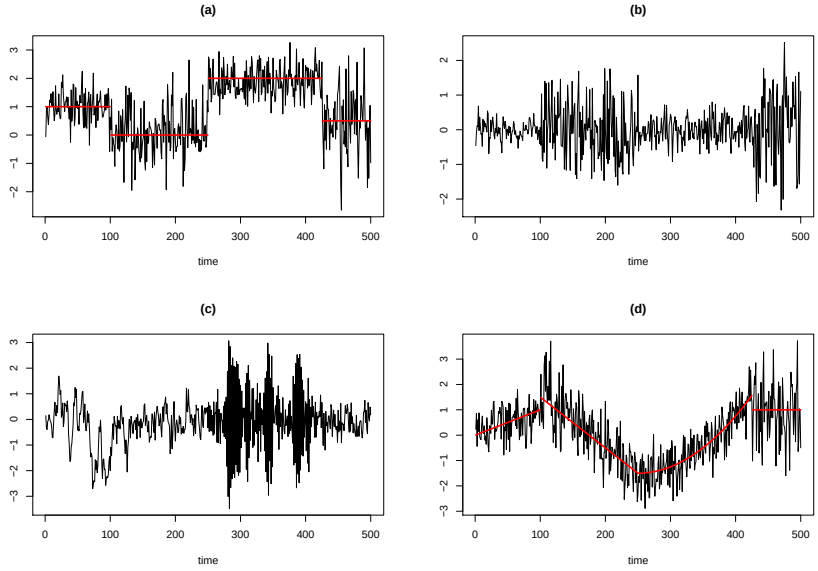

Figure 1.1: Realisations of the piecewise linear regression model. (a) Change in (constant) mean; (b) Change in variance; (c) piecewise AR model; and (d) piecewise quadratic mean. In all cases the changepoints are at time-points 100, 250 and 425. For plots (a) and (d) the underlying mean is shown.

changes in mean related to other distributional forms including gamma [30], exponential [25] and binomial [29]; and also to changes in variance within normally distributed observations by [24] and [10].

Recalling our changepoint problem formulation above, we can construct a test statistic which will decide whether a change has occurred. The likelihood ratio method requires calculating the maximum log-likelihood value under both null and alternative hypotheses. For the null hypothesis the maximum log-likelihood value is just $\log p(y_{1:n}|\hat{\theta})$.

Under the alternative hypothesis, consider a model with a changepoint at τ, with $\tau \in \{1, 2, \ldots, n-1\}$. Then the maximum log likelihood for a given τ (the profile log-likelihood for τ) is

$$Prl(\tau) = \log p(y_{1:\tau}|\hat{\theta}_1) + \log p(y_{(\tau+1):n}|\hat{\theta}_2).$$

The maximum log-likelihood value under the alternative is just $\max_\tau Prl(\tau)$. This results in the test statistic

$$\lambda = 2\left[\max_\tau Prl(\tau) - \log p(y_{1:n}|\hat{\theta})\right].$$

The test involves choosing a threshold, c, such that we reject the null hypothesis if $\lambda > c$. If we reject the null hypothesis, which corresponds to detecting a changepoint, then we estimate its position as $\hat{\tau}$ the value of τ that maximises $Prl(\tau)$.

Note that changepoint problems are not regular, so the usual asymptotic results of the likelihood ratio statistic do not apply. Full derivations of the asymptotic distribution for the likelihood ratio test of univariate and multivariate normal, gamma, binomial and poisson distributions are provided by [11]. These can be used to give an approximate threshold for any required significance level.

The likelihood-ratio framework can naturally extend to detecting changes in a subset of the parameters; for example for the model in Example 1, we may be interested in changes only in the regression parameters, or a specific subset of the regression parameters. Such problems only require a change in the calculation of the maximum likelihood for each model, with maximisation of θ_1 and θ_2 being done over appropriate constraints for the parameters.

1.2.2 Penalised likelihood approaches

The use of penalised likelihood approaches have been popular within the changepoint literature (see for example [23] or [54]). The popularity of this approach stems from parsimony arguments. These methods more naturally extend to the multiple changepoint setting than does the likelihood-ratio statistic approach. Below we outline a general approach for the detection of changepoints using penalised likelihood.

We begin by defining the general penalised likelihood.

Definition 1. *Consider a model $\mathcal{M}_k$, with p_k parameters. Denote the parameters by Θ_k, and the likelihood by $L(\Theta_k)$. The penalised likelihood is defined to be:*

$$PL(\mathcal{M}_k) = -2\log\max L(\Theta_k) + p_k\phi(n),$$

where $\phi(n)$ is the penalisation function, which is an non-decreasing function of the length of the data, n.

The value of $\mathcal{M}_k$ that minimises $PL(\mathcal{M}_k)$ is deemed the most appropriate model. Obviously the choice of model will depend on the choice of penalty function $\phi(n)$. Various penalty functions can be considered, including Akaike's information criterion (AIC) [1], Schwarz information criterion (SIC) [42] and the Hannan-Quinn information criterion [26]. These criteria are defined as follows:

$$\begin{aligned} \text{AIC} &: \phi(n) = 2 \\ \text{SIC} &: \phi(n) = \log n \\ \text{Hannan-Quinn} &: \phi(n) = 2\log\log n. \end{aligned}$$

Whilst the AIC is a popular penalty term, it has been shown that it asymptotically overestimates the correct number of parameters. Thus as the SIC and Hannan-Quinn criteria both asymptotically estimate the correct number of parameters, these are generally preferred. (See [54] for details of the SIC case.)

For the changepoint problem described in Section 1.1.1, $\mathcal{M}_k$ corresponds to the model with k changepoints. The associated parameter vector is $\Theta_k = (\tau_{1:k}, \theta_{1:k+1})$, which has dimension $p_k = k + (k+1)\text{dim}(\theta)$. For detecting a single changepoint the calculation of the two penalised likelihoods corresponding to either one or no changepoint, involves a similar likelihood maximisation step to that described in Section 1.2.1.

For estimating a single changepoint, there is a close correspondence between the penalised likelihood and the likelihood-ratio test approaches. Both involve comparing the maximum log-likelihood of the two models corresponding to one and no changepoint. A changepoint is detected if the increase in log-likelihood under the one changepoint model is greater than some threshold. The differences lie only in how this threshold is calculated.

1.2.3 Bayesian Methods

To perform a Bayesian analysis we need to specify a prior probability for there being a changepoint, $\Pr(M = 1)$, and conditional on there being a changepoint, a distribution for its position $p(\tau)$. Note that $\Pr(M = 0) = 1 - \Pr(M = 1)$.

Firstly consider the case where the hyperparameters ψ are known. In this case it is straightforward to write down the posterior distribution in terms of marginal likelihoods, $Q(s, t)$, as defined in (1.1). The posterior is

$$\begin{aligned} \Pr(M = 0|y_{1:n}) &\propto \Pr(M = 0)Q(1, n; \psi) \\ \Pr(M = 1, \tau|y_{1:n}) &\propto \Pr(M = 1)p(\tau)Q(1, \tau; \psi)Q(\tau + 1, n; \psi), \text{ for } \tau = 1, \ldots, n - 1. \end{aligned}$$

In the case on which we focus, where the marginal likelihoods can be calculated analytically, this posterior is simple to calculate. It requires calculating the above expressions to be evaluated and normalised to give the posterior probabilities. This is an $O(n)$ calculation. As mentioned above, in cases where we can calculate the marginal likelihood, we can normally calculate analytically the conditional posterior for segment parameters given the start and end of the segment. Thus we can extend the above calculation to give the joint posterior of whether there is a changepoint, its position if there is one, and the segment parameters.

If we focus on purely detecting whether there is a changepoint, then we get

$$\frac{\Pr(M = 1|y_{1:n})}{\Pr(M = 0|y_{1:n})} = \frac{\Pr(M = 1)}{\Pr(M = 0)} \left(\frac{\sum_{\tau=1}^{n-1} p(\tau)Q(1, \tau; \psi)Q(\tau + 1, n; \psi)}{Q(1, n; \psi)} \right).$$

The last term on the right-hand side is called the Bayes Factor. Thus the posterior ratio of probabilities of one changepoint to no changepoint is the prior ratio multiplied by the Bayes Factor. As such the Bayes Factor quantifies the evidence in the data for the model with one changepoint, as opposed to the model with no changepoint.

Note that the posterior distribution depends on ψ. In particular the choice of ψ can have considerable effect on the posterior probability for a changepoint. The reason for this is linked to Bartlett's paradox [5], which shows that when comparing nested models, the use of improper priors for the parameters in the more complex model will lead to posterior of probability of one assigned to the simpler model. Even when we do not use improper priors, choices of ψ that correspond to vague priors for the segment parameters will tend to prefer the simpler model, that is inferring no changepoint. We will return to this issue in the simulation study in Section 1.4.

There are two approaches to deal with choosing ψ, in the absence of prior information. The first is to introduce a prior on ψ. In this case we can define the marginal likelihood for ψ as

$$\mathrm{ML}(\psi) = \Pr(M = 0)Q(1, n; \psi) + \sum_{\tau=1}^{n-1} \Pr(M = 1)p(\tau)Q(1, \tau; \psi)Q(\tau + 1, n; \psi),$$

and let $p(\psi)$ denote the prior. Then the marginal posterior for ψ is proportional to $p(\psi)\mathrm{ML}(\psi)$, which could be explored using MCMC. Note that it is possible to choose an improper prior for ψ, as this is a parameter common to both the no changepoint and one changepoint models.

Computationally simpler is to adopt an empirical Bayes approach – and use the data to get a point estimate for ψ. For example, optimisation algorithms can be used to find the value of ψ that maximises $\mathrm{ML}(\psi)$, and then inference can be made

conditional on this value for ψ. This approach has the disadvantage of ignoring the effect of uncertainty in the choice of ψ.

We do not go into detail for either approach here, though we will return to this issue when discussing Bayesian methods for multiple changepoint problems. Also, in Section 1.4 we look empirically at and compare the different approaches for dealing with no knowledge about ψ.

1.3 Multiple Changepoint Models

Many of the ideas for analysing single changepoint models can be adapted, at least in theory, to the analysis of multiple changepoint models. However, the analysis of multiple changepoint models is computationally much more challenging, as the number of possible positions of m changepoints increases quickly with m. For example, with $1,000$ data points there are just 999 possible positions of a single changepoint, but 2×10^{23} sets of possibilities for 10 changepoints. Much of the focus of the following sections is on the resulting computational challenge of detecting multiple changepoints.

We first focus on two general search methods, which can be used to extend the likelihood-ratio statistic approach to detecting multiple changepoints, and can be used to efficiently perform the maximisation required in applying penalised likelihood methods. We then introduce a new criteria for detecting changepoints, based on minimum description length, and show how the latter of these search methods can be used to find the optimal set of changepoints in this case. Finally we describe how to efficiently perform a Bayesian analysis.

1.3.1 Binary Segmentation

The binary segmentation algorithm is perhaps the most established search algorithm used within the changepoint literature. Early applications of the binary segmentation search algorithm include [43] and [44]. For details on the consistency of the binary segmentation approach for estimating the true changepoint locations, $\tau_{1:m}$, under various conditions, the reader is referred to the work of [49] and [48].

Binary segmentation can be used to extend any single changepoint method to multiple changepoints. We begin by initially applying this detection method to the whole data. If no changepoint is detected we stop, otherwise we split the data into two segments (before and after the changepoint), and apply the detection method to each segment. If a changepoint is detected in either, or both, segments, we split these into further segments, and apply the detection method to each new segment. This procedure is repeated until no further changepoints are detected.

Generic pseudo-code for one implementation of this is given in Algorithm 1. This considers a general test statistic $\Lambda(\cdot)$, estimator of changepoint position $\hat{\tau}(\cdot)$, and rejection threshold C. The idea is that the test statistic is a function of data, such as the likelihood ratio statistic, and we detect a changepoint in data $y_{s:t}$ if $\Lambda(y_{s:t}) > C$. If we detect a changepoint, our estimate of its position, such as the maximum likelihood estimate, is $\hat{\tau}(y_{s:t})$. Within the code $\mathcal{C}$ denotes the set of detected changepoints, and $\mathcal{S}$ denotes a set of segments of the data that need to be tested for a changepoint. One iteration chooses a segment from $\mathcal{S}$, and performs the test for a changepoint. For a negative result the segment is removed from $\mathcal{S}$. Otherwise a changepoint is detected and added to $\mathcal{C}$, and the segment is replaced in $\mathcal{S}$ by two segments defined by splitting the original segment at the changepoint. Note in step 3(b), r is just the position of the changepoint in the original data set, calculated from $\hat{\tau}(y_{s:t})$, the position of the changepoint in the segment $[s, t]$.

In steps 3(c) and 3(d) we only add new segments to $\mathcal{S}$ if they contain at least 2 observations: otherwise the new segments can not contain further changepoints.

Algorithm 1 The Generic Binary Segmentation Algorithm to find all possible change points.

Input: A set of data of the form, $(y_1, y_2, \ldots, y_n)$.
A test statistic $\Lambda(\cdot)$ dependent on the data.
An estimator of changepoint position $\hat{\tau}(\cdot)$.
A rejection threshold C.

Initialise: Let $\mathcal{C} = \emptyset$, and $\mathcal{S} = \{[1, n]\}$

Iterate while $\mathcal{S} \neq \emptyset$

1. Choose an element of $\mathcal{S}$; denote this element as $[s, t]$.
2. If $\Lambda(y_{s:t}) < C$, remove $[s, t]$ from $\mathcal{S}$.
3. If $\Lambda(y_{s:t}) \geq C$ then:
   (a) remove $[s, t]$ from $\mathcal{S}$;
   (b) calculate $r = \hat{\tau}(y_{s:t}) + s - 1$, and add r to $\mathcal{C}$;
   (c) if $r \neq s$ add $[s, r]$ to $\mathcal{S}$;
   (d) if $r \neq t - 1$ add $[r + 1, t]$ to $\mathcal{S}$.

Output the set of change points recorded $\mathcal{C}$.

Binary segmentation is a fast algorithm, that can be implemented with computational cost $O(n)$ where n is the length of data. However, it can be difficult to choose C appropriately – and different choices of C can lead to substantial differences in the estimate of the number of changepoints. An alternative approach to detecting multiple changepoints by recursively applying a single changepoint method is given in [31].

1.3.2 Segment Neighbourhood Search

[7] and [6] consider an alternative search algorithm for changepoint detection, namely the Segment Neighbourhood approach (also referred to as Global Segmentation). The basic principle of this approach is to define some measure of data fit, $R(\cdot)$ for a segment. For inference via penalised likelihood we would set $R(y_{s:t})$ to be minus the maximum log-likelihood value for data $y_{s:t}$ given it comes from a single segment. That is

$$R(y_{s:t}) = -\log p(y_{s:t}|\hat{\theta}). \tag{1.4}$$

We then set a maximum number of segments, M, corresponding to at most $M - 1$ changepoints.

The segment neighbourhood search then uses a dynamic programming algorithm to find the best partition of the data into $m + 1$ segments for $m = 0, \ldots, M - 1$. The best partition is found by minimising the cost function $\sum_{i=0}^{m} R(y_{\tau_i:\tau_{i+1}})$ for a partition with changepoints at positions $\tau_1, \tau_2, \ldots, \tau_m$. Thus for $R(\cdot)$ defined in (1.4), this would give the partition of the data with m changepoints that maximises the log-likelihood. The algorithm will output the best partition for $m = 0, \ldots, M - 1$, and the corresponding minimum value of the cost function, which we denote $c_{1,n}^m$.

For the choice of $R(\cdot)$ given by (1.4), $2c_{1,n}^m$ will be minus twice the log-likelihood. So choosing m based on penalised likelihood is achieved by choosing m to minimise $2c_{1,n}^m + p_m\phi(n)$; where p_m is the number of parameters in the model with m change-

points, and $\phi(n)$ is the penalty function. The best partition found by the algorithm for that value of m gives the positions of the detected changepoints.

Generic pseudo-code for this approach can be seen in Algorithm 2, and is based on a dynamic programming approach described by [2]. The drawback of this approach is it's computational cost. The segment neighbourhood search is an $O(n^2)$ computation; compared with $O(n)$ for the binary segmentation algorithm. However this cost does result in improved predictive performance in simulation studies [6].

Algorithm 2 The Generic Segment Neighbourhood Algorithm to find up to $R-1$ change points.

Input: A set of data of the form, $(y_1, y_2, \ldots, y_n)$.
A measure of fit $R(\cdot)$ dependent on the data which needs to be minimised.
An integer, $M-1$ specifying the maximum number of change points to find.

Initialise: Let $n =$ length of data.
Calculate $q^1_{i,j} = R(y_{i:j})$ for all $i, j \in [1, n]$ such that $i < j$.

Iterate for $m = 2, \ldots, M$

1. Iterate step 2 for all $j \in \{1, 2, \ldots, n\}$.
2. Calculate $q^m_{1,j} = \min_{v \in [1,j)} (q^{m-1}_{1,v} + q^1_{v+1,j})$.
3. Set $\tau_{m,1}$ to be the v that minimises $(q^{m-1}_{1,v} + q^1_{v+1,n})$.
4. Iterate step 5 for all $i \in \{2, 3, \ldots, M\}$.
5. Let $\tau_{m,i}$ to be the v that minimises $(q^{m-i-1}_{1,v} + q^1_{v+1,cp_{m,i-1}})$.

Output For $m = 1, \ldots, M$: the total measure of fit, $q^m_{1,n}$ for $m-1$ change points and the location of the change points for that fit, $\tau_{m,1:m}$.

1.3.3 Minimum Description Length

[12] propose the use of the minimum description length (MDL) principle to estimating changepoints. The basic idea is that the best fitting model is one which enables maximum compression of the data. For a given set of changepoints we can estimate what is called the code-length of the data. Loosely, this code length is the amount of memory space needed to store that data. We thus estimate the number and position of the changepoints as the set of changepoints which have the minimum code-length. See [12] and references therein for further background to MDL.

Our aim here is to show how finding the best set of changepoints under MDL can be achieved using the segment neighbourhood algorithm. This guarantees finding the optimal set of changepoints according to the MDL criterion. By comparison, [12] use a complicated genetic algorithm to fit the model.

For concreteness we will focus on the model of Section 1.1.2. In this case, up to proportionality, the code-length for a set of m changepoints, $\tau_1, \ldots, \tau_m$ is defined as

$$\mathcal{CL}(m; \tau_{1:n}) = -\sum_{i=1}^{m+1} \log p(y_{(\tau_{i-1}+1):\tau_i}|\hat{\theta}_i) + \log(m+1) + (m+1)\log(n) + \sum_{i=1}^{m+1} \frac{d+1}{2} \log n_i,$$

where $n_i = \tau_i - \tau_{i-1}$ is the length of segment i, and $d+1$ is the dimension of the parameter vector associated with each segment. (See [12] for the derivation.)

Now denote $R(y_{s:t}) = -\log p(y_{s:t}|\hat{\theta}) + \frac{d+1}{2}\log(t-s+1)$. We can re-write the

code-length as

$$\mathcal{CL}(m;\tau_{1:n}) = \sum_{i=1}^{m+1} R(y_{(\tau_{i-1}+1):\tau_i}) + \log(m+1) + (m+1)\log(n).$$

Thus we can use the segment neighbourhood algorithm to calculate

$$c_{1,n}^m = \min_{\tau_{1:m}} \sum_{i=1}^{m+1} R(y_{(\tau_{i-1}+1):\tau_i}),$$

for $m = 0, \ldots, M-1$. We then estimate the number of changepoints as the value m which minimises $c_{1,n}^m + \log(m+1) + (m+1)\log(n)$. The segment neighbourhood algorithm also outputs the optimal set of changepoints.

1.3.4 Bayesian Methods

For a Bayesian analysis we need to specify a prior for the number and position of changepoints. There are two approaches. The first is to specify a prior on the number of changepoints, and then a prior for their position given the number of changepoints [22]. The second is to specify the prior for the number and position of changepoints indirectly through a distribution for the length of each segment. The latter has computational advantages [17] and is more natural in many applications. For example it means that the prior does not need to be adapted based on the period of time over which the time-series is observed. It is also easier to use inferences from similar data sets, which maybe of different length, to construct appropriate priors. We thus focus solely on this form of prior.

Formally we introduce a probability mass function, denoted $g(\cdot;\psi)$, to be the mass function for the length of a segment. We allow there to be unknown parameters of this mass function, and these will be part of the hyperparameters of the model: hence the dependence on ψ. Associated with the mass function will be a survivor function $S(\cdot;\psi)$, which satisfies $S(t;\psi) = \sum_{i=t}^{\infty} g(i;\psi)$.

With this construction, the prior probability for m changepoints at positions $\tau_1, \ldots, \tau_m$ will be

$$p(m, \tau_{1:m}|\psi) = \left(\prod_{i=1}^{m} g(\tau_i - \tau_{i-1};\psi)\right) S(\tau_{m+1} - \tau_m;\psi),$$

where as before we set $\tau_0 = 0$ and $\tau_{m+1} = n$. This prior corresponds to a product-partition model [3, 4].

A common choice for the distribution of the segment lengths is the geometric distribution with parameter p. In this case $g(t;\psi) = p(1-p)^{t-1}$, $S(t;\psi) = (1-p)^{t-1}$, and $p(m, \tau_{1:m}|\psi) = p^m(1-p)^{n-m-1}$. Note that this corresponds to a binomial prior on the number of changepoints, and a conditional uniform prior on their position.

We now derive the posterior conditional on a fixed value of ψ. Under the assumption that we can calculate the segment marginal likelihoods (1.1), we can integrate out the parameters associated with each segment to obtain the following marginal posterior for the number and position of changepoints

$$\begin{aligned} p(m, \tau_{1:m}|\psi, y_{1:n}) \quad \propto \quad & \left(\prod_{i=1}^{m} g(\tau_i - \tau_{i-1};\psi) Q(\tau_{i-1}+1, \tau_i;\psi)\right) \\ & \times S(\tau_{m+1} - \tau_m;\psi) Q(\tau_m + 1, \tau_{m+1};\psi). \end{aligned} \tag{1.5}$$

The normalising constant is just the marginal likelihood for ψ. As mentioned above, for models where we can calculate the segment marginal likelihoods we can usually simulate from the posterior distribution of the segment parameters given the changepoint positions. Thus if we can generate samples from this posterior on the number and position of the changepoints, it is straightforward to sample from the joint posterior of the changepoints and the segment parameters. While MCMC [36] and reversible jump MCMC methods [22] can be used to generate (approximate) samples from the posterior (1.5). These methods can be computationally intensive, and lead to difficulties of diagnosing convergence of the MCMC algorithm. For example the analysis of the coal-mining disaster data in [22] is incorrect due to the MCMC algorithm not being run for long enough [17].

Instead, we describe a computationally efficient algorithm that can generate iid samples from this posterior. The algorithm we describe has two stages. The first is a forward pass through the data; the second involves simulating the changepoints backwards in time. The algorithm is thus related to the forward-backward algorithm for hidden Markov models [18]. However the basic idea underlying this approach dates back to work by [53]; see also the methods of [3] and [38]. The version we give is suitable for online analysis of data.

For this algorithm we introduce a variable C_t to be the position of the most recent changepoint prior to time t. Thus $C_t \in \{0, 1, \ldots, t-1\}$, with $C_t = 0$ denoting no changepoint prior to t. Note that either $C_t = t-1$, or $C_t = C_{t-1}$, depending on whether or not there is a changepoint at time $t-1$. The forward algorithm calculates $\Pr(C_t = i | y_{1:t}, \psi)$ for $i = 0, \ldots, t-1$. It is based on the following recursion. For $t = 2, \ldots, n$ we have

$$\Pr(C_t = i | y_{1:t}, \psi) \propto \Pr(C_{t-1} = i | y_{1:t-1}, \psi) \left(\frac{S(t-i; \psi)}{S(t-i-1; \psi)} \right) \left(\frac{Q(i+1, t; \psi)}{Q(i+1, t-1; \psi)} \right), \tag{1.6}$$

for $i = 0, \ldots, t-2$; and

$$\Pr(C_t = t-1 | y_{1:t}, \psi) \propto Q(t, t; \psi) \sum_{j=0}^{t-2} \Pr(C_{t-1} = j | y_{1:t-1}, \psi) \left(\frac{g(t-j-1; \psi)}{S(t-j-1; \psi)} \right). \tag{1.7}$$

Recursion (1.6) corresponds to no changepoint at time $t-1$. Thus $C_t = C_{t-1}$ and hence the final two terms correspond to the prior probability of this, and the likelihood of the new observation given $C_t = i$ respectively. Recursion (1.7) corresponds to a changepoint at time $t-1$. In which case $Q(t, t; \psi)$ is the likelihood of the observation, and the sum is the probability of a changepoint at $t-1$ prior to observing y_t. These recursions are initiated with $\Pr(C_1 = 0 | y_1) = 1$. For more details of the derivation see [19]. Details of how the output from these recursions can be used to calculate the marginal likelihood for ψ are given in [18].

The backward step generates samples from the posterior of the position and number of changepoints. It requires that the probabilities $\Pr(C_t = i | y_{1:t})$ have been stored for all $t = 1, \ldots, n$ and $i = 0, \ldots, t-1$. To generate one sample of changepoints we first simulate the last changepoint from the distribution of C_n given $y_{1:n}$. Denote the changepoint position by t. Then if $t > 0$ we can simulate the next changepoint back in time, C_t, from the conditional distribution

$$Pr(C_t = i | y_{1:n}, C_{t+1} = t, \psi) \propto \Pr(C_t = i | y_{1:t}, \psi) \left(\frac{g(t-i; \psi)}{S(t-i; \psi)} \right), \text{ for } i = 1, \ldots, t-1.$$

(1.8)

(Note the event $C_{t+1} = t$ just corresponds to there being a changepoint at t.) The calculation of this probability mass function uses the fact that conditional on a changepoint at t, the data after this changepoint is independent of the changepoints before t. We recursively simulate changepoints backwards in time from (1.8) until we first simulate $C_t = 0$.

Algorithm 3 Algorithm for simulating from the posterior distribution of changepoint positions.

Input: A set of data of the form, $(y_1, y_2, \ldots, y_n)$.
A value for the hyperparameters ψ.
Survivor functions for segment lengths $S(\cdot; \psi)$.
A weight function $W(\cdot; \psi)$, such that $W(y_{s:t}; \psi) = Q(y_{s:t}; \psi)/Q(y_{s:t-1}; \psi)$ for $t > s$, and $W(y_s; \psi) = Q(y_s; \psi)$; where $Q(\cdot; \psi)$ is defined in (1.1).
The number of samples from the posterior, N.

Initialise: Let $t = 2$. Let $\gamma_0^{(1)} = 1$.

Iterate for $t = 2, \ldots, n$

1. For $i = 0, \ldots, t-2$; set
$$\gamma_i^{(t)} = \gamma_i^{(t-1)} \left(\frac{S(t-i; \psi)}{S(t-i-1; \psi)} \right) W(y_{i+1:t}; \psi).$$
2. Set
$$\gamma_{t-1}^{(t)} = W(y_t; \psi) \sum_{j=0}^{t-2} \gamma_j^{(t-1)} \left(\frac{S(t-j-1; \psi) - S(t-j; \psi)}{S(t-j-1; \psi)} \right) ..$$
3. Normalise $\gamma^{(t)}$s. Set $A = \sum_{i=0}^{t-1} \gamma_i^{(t)}$, and for $i = 0, \ldots, t-1$ set $\gamma_i^{(t)} = \gamma_i^{(t)}/A$.

Iterate for $j = 1, \ldots, N$

1. Simulate from the distribution with mass $\gamma_i^{(n)}$ associated with realisation i for $i = 0, \ldots, n-1$; denote the realisation by t.
2. If $t > 0$, set $\mathcal{C}_j = \{t\}$; otherwise $\mathcal{C}_j = \emptyset$.
3. While $t > 0$ repeat steps 4 and 5.
4. Simulate from the distribution with mass proportional to
$$\gamma_i^{(t)} \left(\frac{S(t-i; \psi) - S(t-i+1; \psi)}{S(t-i; \psi)} \right),$$
associated with realisation i for $i = 0, \ldots, t-1$; denote the realisation by t.
5. If $t > 0$, update $\mathcal{C}_j = \{t, \mathcal{C}_j\}$

Output the N sets of changepoints, $\mathcal{C}_1, \ldots, \mathcal{C}_N$.

Full details of the forward recursion and backward simulation algorithm are given in Algorithm 3. In this algorithm $\gamma_i^{(t)}$ denotes $\Pr(C_t = i | y_{1:t})$.

The algorithm has a computational and storage cost that is quadratic in n, the number of data points. This is because the support of C_t increases linearly with t. However, for large t, the majority of the probabilities $\Pr(C_t = i | y_{1:t})$ are negligible. Hence computational and storage savings can be made by pruning such probabilities. See [19] for a principled way of implementing such pruning, which results in an algorithm with computational and storage costs that are $O(n)$. Pruning does introduce approximation error, but empirical results [19] suggest these approximations are negligible. The resulting algorithms can analyse large data sets efficiently, see [19] and [20] for applications to genomic data. Even in these applications, where n is of the order of tens of thousands, and there may be hundreds of changepoints, generating thousands of samples from the posterior takes a matter

of seconds.

Thus we have a simple, efficient and accurate method for Bayesian inference in the case that the hyperparameters, ψ, are known. In cases where this is not true, we can either introduce a prior on ψ or estimate ψ from the data. The former is the fully-Bayesian approach, but comes at a computational cost. Inference will require the use of MCMC, or related techniques. The above algorithm can be used within MCMC to help mixing. However, this can be computationally expensive – as the forward recursions will need to be solved for each proposed new value for ψ. (See [17] for discussion of this and suggestions for efficiently implementing an MCMC approach.) The alternative is to estimate ψ from the data – for example through maximising the marginal likelihood. Performing the maximisation is often possible via a Monte Carlo Expectation Maximisation (EM) algorithm [50]. Results in [16] suggest that such an approach loses little in terms of statistical efficiency, but is computationally more efficient than the fully Bayesian solution of introducing a prior on ψ.

1.4 Comparison of Methods

We now compare different changepoint methods for the problem of detecting a change in variance. In general detecting changes in variance is more challenging than detecting changes in mean, and is important in applications such as finance and environmental applications [34]. Compared to the change in mean problem, [10] observe the detection of changes in variance has received comparatively little attention. We will look in turn at the problem of detecting a single changepoint and multiple changepoints.

1.4.1 Single Changepoint Model

We first present a simulation study which aims to compare the frequentist and Bayesian methods for detecting a single changepoint, and to look at how specification of the hyperparameter ψ can affect the Bayesian inferences. We base our study on a specific case of the model described in Example 1. Each data point has a normal distribution with mean 0, but we allow for the possibility of the variance changing at a changepoint. Details of the analytic calculations of maximum likelihood estimates, posterior distributions and marginal likelihoods for the segments are given in the Appendix.

In particular we simulated time-series consisting of 200 observations. For the first 100 data points, the observations were iid from a standard normal distribution. The second 100 data points were iid from a normal distribution with mean 0 and variance σ^2. We considered 6 scenarios, each with different values of σ: $\sigma^2 = 1$, 1.25, 1.5, 2, 3 and 4. The first scenario corresponds to no changepoint, as the distribution of the data is identical for all 200 data points and is used to estimate false-positive rates for different methods. The remaining scenarios correspond to increasingly large changes. We simulated 10,000 independent data sets for each scenario.

Comparison of Method

We first looked at the performance of various methods to detect a changepoint within a series. For detecting a changepoint, each method is based upon comparing a statistic, such as the Bayes Factor or the likelihood ratio statistic, with a threshold value. The threshold value will affect both the false-positive rate and also the

proportion of true changepoints (true-positives) detected for a given value of σ^2. By varying this threshold we can plot how the latter varies with the former, and we give the results in a so-called receiver operating characteristic (ROC) curve. This enables us to calibrate the comparison of methods, so we compare the true-positive rate of different methods for a common false-positive rate.

For the Bayesian implementation the hyperparameters, ψ, are the parameters of the inverse-gamma distribution for the segment variance. Initially we set the shape parameter to be 2, and the scale parameter so that the mean of the distribution was the sample variance of the data. The results, in terms of the ROC curve, were robust to these choices; but we investigate below the effect of the choice of ψ on the performance of the Bayesian approach.

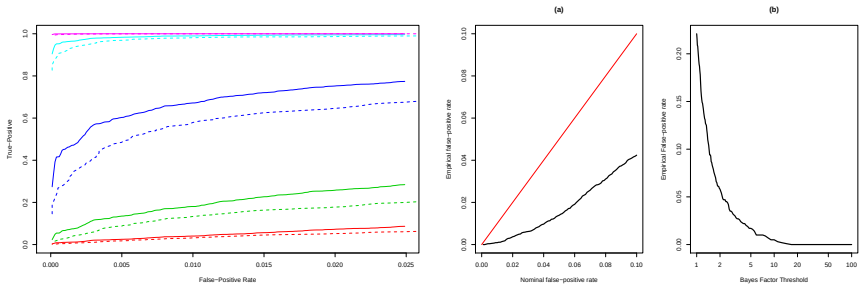

Figure 1.2: (a) ROC curves for the Bayesian (full-lines) and frequentist (dashed-lines) approaches. Each pair of lines corresponds to a different value of σ, from bottom to top: 1.25, 1.5, 2, 3 and 4. (b) Nominal false-positive rate versus empirical false-positive rate for the likelihood-ration method. (c) Bayes Factor threshold versus empirical false-positive rate for Bayesian method.

Results are shown in Figure 1.2(a). Both the likelihood ratio and penalised likelihood methods (where we vary the penalty) give identical ROC curves, see Section 1.2.2, so we plot a single curve for both these. The results show similar performance for the Bayesian and frequentist approaches, with the Bayesian method having slightly greater power, particularly for intermediate values of σ. The intuition behind this is that for detecting change in variance there is normally substantial uncertainty about the position of the changepoint. The Bayes factor averages over this uncertainty, so allows for the accumulation of evidence for a changepoint; where as frequentist methods depend only on the fit for the most likely changepoint position – and as such ignores any information from other possible changepoint locations.

Implementation of Methods

The comparison above looks at overall performance of methods via an ROC curve, which look at false and true positive rates for a range of threshold values for each method. However, when implementing a method in practice we need guidelines for choosing this threshold.

For the likelihood-ratio approach, there is clear guidance on choosing the threshold based on asymptotic results which give nominal false-positive rates for different threshold values [11]. In Figure 1.2(b) we plot empirical false-positive rates for a range of nominal false-positive rates. For the size of data we analysed, the nominal false-positive rates over-estimate the true false positive-rates, typically by a factor of around 2.

For comparison, we calculated the false-positive rates for the three penalised

likelihood methods introduced in Section 1.2.2. These are AIC, SIC and Hannan-Quinn. For our data $n = 200$ so $\phi(n) = 2$, 5.3 and 3.3 respectively. The false-positive rates were 70%, 4.4% and 26% in turn. In particular this suggests that the penalty used in AIC is too small, and results in over-detection of changepoints.

For the Bayesian approach, the test is affected by (i) the prior probability of a changepoint; (ii) a threshold on the posterior probability for detecting a changepoint; and (iii) the choice of ψ. Strictly (ii) should be specified by considering the relative cost of falsely detecting a changepoint to missing one. The larger this is, the higher the threshold. However, in many cases it can be difficult to specify this, and also often there is little prior information to guide (i). In these cases, it is common to use general rules of thumb for the Bayes factor [33].

In practice, the most important choice is (iii), the prior for ψ. Furthermore it can be hard to predict the effect that this choice will have on the properties of the test. In particular we want to guard against choosing values of ψ that correspond to weakly informative priors which will lead to preference for the model for no changepoint.

To investigate the effect of the choice of ψ we repeated the above analysis but for a range of values for ψ, the parameters of the inverse gamma distribution for the variance. In each case we chose parameters so that the mean of the inverse gamma distribution was equal to the empirical mean, and just considered choice of the shape parameter, a. The choice $a \approx 0$ corresponds to a weakly informative prior. Results are given in Figure 1.3(a). We observe that small and large values of a lead to the detection of a changepoint in fewer data sets. For the Bayesian method to detect changepoints well we need a value of a that leads to a prior distribution that is roughly consistent with the variation in σ across the two segments.

As discussed in Section 1.2.3, the two approaches to choosing ψ based on the data are to introduce a hyperprior on ψ or an empirical Bayes approach of estimating ψ by maximising the marginal likelihood. We tried both approaches. They provided almost identical results, so here we give the results for the empirical Bayes approach. For a threshold value of 10 for the Bayes factor for the model with no changepoints against the model with one, the false positive rate was 0.005, with, for increasing values of σ, true-positive rates of 0.02, 0.13, 0.63, 0.98 and 1.0.

For this approach we looked at how the empirical false-positive rate varies with the threshold used for the Bayes Factor. This is shown in 1.2(c). Note that it is difficult to predict the form of the relationship beforehand. For this example, a threshold of around 2, corresponding to twice as much evidence for one changepoint as opposed to no changepoints, corresponds to a false positive rate of 5%. Note also that a threshold of 1, which corresponds to equal evidence in the data for either one changepoint or no changepoints, has a false-positive rate much lower than 0.5, which is what we may have predicted.

1.4.2 Multiple Changepoint Model

We now consider analysis of multiple changepoint models. We aim to look at the relative performance of the different methods and to quantify what affects the power to detect changepoints.

As in the single changepoint case, we simulated data under a model where the variance changes across segments. We simulated time-series consisting of 2,000 data points. Each data set contained 10 changepoints, which were uniformly distributed subject to the constraint that each segment contained at least 40 observations. Within each segment, the observations were iid draws from a normal distribution with mean 0 and common variance. The distribution of the segment variances were

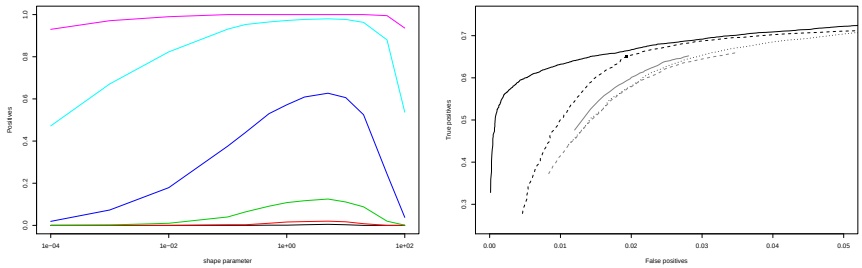

Figure 1.3: (a) Proportion of data sets with Bayes factor for no changepoint > 10, as a function of a, the shape parameter of the inverse Gamma distribution. Each lines corresponds to a different value of σ, bottom to top: 1.0, 1.25, 1.5, 2, 3 and 4. (b) ROC curve for multiple changepoint methods. Bayesian method (black full line); binary segmentation based on likelihood-ratio test (black dotted line); binary segmentation using Bayes Factor (grey dashed line); the [31] approach for segmentation based on the likelihood-ratio test (grey full line); and penalised likelihood (black dashed line). The square dot corresponds to MDL.

log-normal, and the parameters of the log-normal distribution chosen so that 95% of variances lay within the interval $[1/10, 10]$. We simulated 1,000 independent data sets.

The distribution of segment variances was specifically chosen to be different from the inverse-gamma distribution used by the Bayesian method. Also, when implementing the Bayesian approach we assumed a geometric distribution of segment lengths and thus did not use the information that all segments contained at least 40 observations. This avoids any bias towards the Bayesian approach through simulating data from exactly the same class of models that the data is analysed under.

When implementing the Bayesian method we used an empirical Bayes approach, estimating hyper-parameters based on maximising the marginal likelihood. The marginal likelihood was maximised using a Monte Carlo EM algorithm.

Comparison of Methods

Firstly we compared different methods based on ROC curves. Making a comparison is non-trivial as the output of Bayesian and frequentist approaches differ. The former gives posterior probabilities for changepoints at each location, while the latter return a list of inferred changepoint positions. The following approach was used, which gives comparison between false and true positive rates for both methods.

For the Bayesian approach we counted a changepoint as detected if the posterior probability of a changepoint within a distance of 20 time-points either side of the true position was greater than a pre-specified threshold. For false positives we considered non-overlapping windows of similar size that did not contain a true changepoint. A false-positive related to a window for which the posterior probability of a changepoint was above the threshold. For the frequentist methods we used a similar approach. Changepoints were considered detected if we inferred a changepoint with a distance of 20 time-points of the true position. We then considered the same non-overlapping windows which did not contain a changepoint, counting a false positive for every window in which we inferred a changepoint. The false-positive rate thus estimates the probability that we estimate there is a changepoint within a randomly chosen window that contains no changepoint.

Results are given in Figure 1.3(b). We compared the Bayesian approach with a number of frequentist methods. The latter included penalised likelihood and MDL using the segment neighbourhood algorithm, and binary segmentation using the likelihood-ratio test. We also implemented binary segmentation with a test based on Bayes factors [52], and the alternative segmentation strategy of [31], implemented with the likelihood ratio test.

There are a number of features of the results that stand out. Firstly, the uniformly most powerful approach is the full-Bayesian method. This approach performs particularly well for small false-positive rates. Secondly, jointly estimating the changepoints, as in the full-Bayesian method or the penalised likelihood approach, performs better than recursively applying single changepoint detection methods using binary segmentation or the approach of [31]. This supports the results of [6].

Thirdly of the two approaches for recursively applying single changepoint methods, that of [31] performed better than binary segmentation. This is perhaps a little surprising, as this method is used much less in the literature. Finally we notice that although the Bayesian method performed better in the single changepoint simulation study, there is very little difference between the binary segmentation approach that used likelihood ratio and the one that used Bayes Factors.

While most approaches can be implemented to give ROC curves, MDL results in a single pair of false-positive and false-negative rates. This pair lies on the penalised likelihood line, and corresponds very closely to the results for penalised likelihood using SIC. Intuitively, this similarity is not surprising as the minimisation criteria for MDL and SIC are very similar (see Sections 1.2.2 and 1.3.3). We also note that using the AIC criteria performed very poorly, detecting over 50 changepoints for each data set. This suggests that the AIC penalty is not large enough.

Factors affecting power

Finally we investigated which factors affect the ability to detect a changepoint, and how this varies across methods. We considered two possible factors, firstly the change in variance and secondly the size of segments either side of the changepoint.

Not surprisingly, the former has an important effect on the ability to detect changepoints. In Figure 1.4(a) we plot, for each changepoint, the posterior probability of a changepoint against the factor by which the variance changes across the changepoint. The former is again calculated by looking for a changepoint within a window which contains all locations a distance of 20 or less from the changepoint position. A change in variance by a factor of 2 has an average posterior probability of about 0.5. While for changes by a factor of 5 or more results in posterior probabilities that are very close to 1.

We then compared power at detecting a changepoint against change in variance. To make the comparison fair, for the Bayesian approach we detect a changepoint if the posterior probability within a window is greater than a threshold. Results for the Bayesian method, MDL and binary segmentation using the likelihood ratio test are compared in Figure 1.4(b). The threshold for the Bayesian approach and for the likelihood ratio test were chosen so that both methods had similar false-positive rates to MDL. The Bayesian approach and MDL have similar power curves, but with evidence that the Bayesian method does better at detecting changepoints when the variance changes by a factor of between 2 and 5. The binary segmentation approach does substantially worse than the other two methods for changepoints across which the variance changes a by factor of 3 or more.

The size of segment had little effect on the probability of detection of a changepoint. Correlation of segment sizes against posterior probability of a changepoint

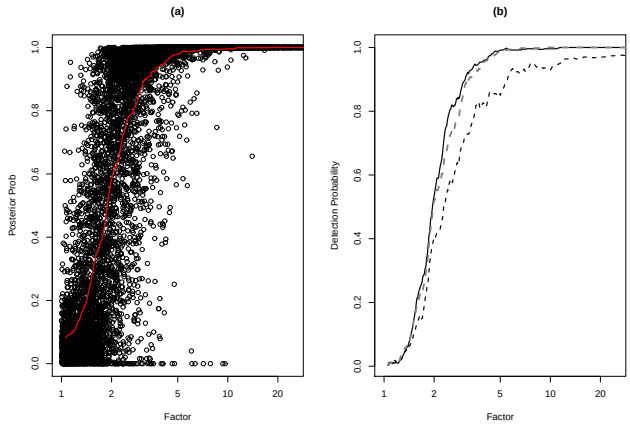

Figure 1.4: (a) Plot of posterior probability of a changepoint against the factor by which the variance changes across the changepoint for each changepoint. A smoothed estimate is given by the line. (b) Plot of power of detecting a changepoint against the factor by which variance changes: Bayesian approach (black full line); MDL (grey dashed line); and binary segmentation with likelihood ratio test (black dashed line).

was around 5%. Similarly small correlation between segment size and detection of changepoints were found for the non-Bayesian methods.

1.5 Conclusion

We have reviewed a number of ways of detecting changepoints, comparing their performance on the problem of detecting changes in variance in a time-series. Analysis of changepoint models is a large area of research, and we have not been able to cover all methods for analysing such models. Examples of alternative approaches include non-parametric methods [39, 40] and methods for online detection based on decision theory [47, 13].

The simulation result suggests that Bayesian methods are the most suitable for this application. One aspect of a Bayesian analysis that we have not reflected on is that the output is a distribution over the number and position of the changepoints. Thus Bayesian methods have the advantage of more easily quantifying uncertainty in changepoint positions than alternative methods. Furthermore, if interest lies in estimating the underlying segment parameters (e.g. how the variance changes over time), a Bayesian approach naturally enables the uncertainty in the changepoints to be taken into account. One disadvantage is that it is harder to summarise or represent the posterior distribution, as compared to methods which output a set of predicted changepoints. One approach is to calculate the most likely (so-called MAP) set of changepoints, which can often be calculated efficiently [9, 16]. However even here there are alternative ways of defining the MAP set of changepoints which can give different results in practice [16].

The main issue when implementing a Bayesian analysis is the choice of priors. For the models we consider here a computationally convenient, yet accurate approach, is to estimate hyperparameters of the prior distributions by maximising the marginal likelihood. This approach appears particularly suitable to multiple changepoint models where there can be substantial information about the hyperparameters due to the variation in parameters across the multiple segments.

When analysing multiple changepoint models, there are computational considerations related to searching for the best set of changepoints or exploring the posterior distribution. For the class of models we focussed on, both of these can be done exactly using either the segment neighbourhood algorithm of Section 1.3.2; or the forward-backward algorithm of Section 1.3.4. Simulation results suggest that using these approaches results in better detection of changepoints than using approximate methods such as binary segmentation. Whilst a complicated genetic algorithm is used to detect changepoints using MDL in [12], we showed that the segment neighbourhood algorithm can be applied for this criteria.

One disadvantage of both the segment neighbourhood algorithm and the forward-backward algorithm is that their computational cost is $O(n^2)$. Approximations to the latter have been suggested in [19], which results in an accurate algorithm whose cost is $O(n)$. One profitable area of future research would be to construct a similar approximate version of the segment neighbourhood algorithm with $O(n)$ computational cost. This is particularly important for applying this approach to analysing the large data sets, such as those currently being analysed in bioinformatics.

A further disadvantage of these two algorithms is that the rely on nice properties of the model. Changepoint models which have strong dependence across segments cannot be analysed by either of these two algorithms. In this case alternatives, such as binary segmentation, MCMC or genetic algorithms, would need to be used to fit models. However, our recommendation is that for models with the appropriate independence properties that these two approaches should be the method of choice for fitting changepoint models.

Acknowledgements

Rebecca Killick is funded by the EPSRC and Shell Research Ltd.

Appendix

Here we give details for estimating segment parameters, conditional on the start and end of the segment, for change in variance model used in the simulation study.

Assume throughout that the segment consists of observations $y_{s:t} = (y_s, \ldots, y_t)$, for $t > s$. There is a single segment parameter, the variance, which we will denote by σ^2. The model assumes that within the segment we have conditionally independent observations with $y_i|\sigma^2 \sim \mathcal{N}(0, \sigma^2)$, for $i = s, \ldots, t$. The maximum likelihood estimator of the parameter is $\hat{\sigma}^2 = \frac{1}{t-s+1}\sum_{i=s}^{t} y_i^2$. The result maximum log-likelihood value is $p(y_{s:t}|\hat{\theta}) = -\frac{n}{2}\left\{\log(2\pi) - \log\hat{\sigma}^2 - 1\right\}$.

For the Bayesian analysis, we have an inverse-gamma prior for σ^2 with hyperparameters $\psi = (a, b)$. The posterior distribution is

$$\sigma^2|y_{s:t} \sim \mathcal{IG}\left(a + \frac{(t-s+1)}{2}, b + \frac{1}{2}\sum_{i=s}^{t} y_i^2\right),$$

with marginal likelihood

$$Q(s,t;\psi) = (2\pi)^{(t-s+1)/2}\frac{\Gamma(a + (t-s+1)/2)b^a}{\Gamma(a)(b + \frac{1}{2}\sum_{i=s}^{t} y_i^2)^{a+(t-s+1)/2)}}.$$

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716 – 723, 1974.

[2] I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54, 1989.

[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 for change point problems. *Journal of the American Statistical Association*, 88:309–319, 1993.

[5] M. S. Bartlett. A comment on D.V.Lindleys statistical paradox. *Biometrika*, 44:533–534, 1957.

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

[7] J. V. Braun and H. G. Muller. Statistical methods for DNA sequence segmentation. *Statistical Science*, 13(2):142–162, 1998.

[8] E. Carlstein, H. G. Muller, and D. Siegmund, editors. *Change-point problems*. Institute of Mathematical Statistics Lecture Notes, 1994.

[9] T. Cemgil, H. J. Kappen, and D. Barber. A generative model for music transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 14:679–694, 2006.

[10] J. Chen and A. K. Gupta. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92:739 – 747, 1997.

[11] J. Chen and A. K. Gupta. *Parametric statistical change point analysis*. Birkhauser, 2000.

[12] R. A Davis, T. C. M Lee, and G. A. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

[13] S. Dayanik, C. Goulding, and H. V. Poor. Bayesian sequential change diagnosis. *Mathematics of Operations Research*, 33:475–496, 2008.

[14] J. B. Elsner, F. N. Xu, and T. H. Jagger. Detecting shifts in hurricane rates using a Markov chain Monte Carlo approach. *Journal of Climate*, 17:2652–2666, 2004.

[15] C. Erdman and J. W. Emerson. A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24(19):2143–2148, 2008.

[16] P. Fearnhead. Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, 53:2160–2166, 2005.

[17] P. Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213, 2006.

[18] P. Fearnhead. Computational methods for complex stochastic systems: A review of some alternatives to MCMC. *Statistics and Computing*, 18:151–171, 2008.

[19] P. Fearnhead and Z. Liu. Online inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B*, 69:589–605, 2007.

[20] P. Fearnhead and D. Vasilieou. Bayesian analysis of isochores. *Journal of the American Statistical Association*, 485:132–141, 2009.

[21] P. Fryzlewicz and S. Subba Rao. Basta: consistent multiscale multiple change-point detection for piecewise-stationary arch processes. *(In submission)*, 2009.

[22] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[23] A. K. Gupta and J. Chen. Detecting changes of mean in multidimensional normal sequences with applications to literature and geology. *Computational Statistics*, 11:211–221, 1996.

[24] A. K. Gupta and J. Tang. On testing homogeneity of variances for gaussian models. *Journal of Statistical Computation and Simulation*, 27:155–173, 1987.

[25] P. Haccou, E. Meelis, and S. Geer. The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic Processes and Their Applications*, 27:121–139, 1988.

[26] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, 41(2):190–195, 1979.

[27] R. Henderson and J. N. S. Matthews. An investigation of changepoints in the annual number of cases of haemolytic uraemic syndrome. *Applied Statistics*, 42:461–471, 1993.

[28] D. V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57:1–17, 1970.

[29] D. V. Hinkley and E. A. Hinkley. Inference about the change-point in a sequence of binomial random variables. *Biometrika*, 57:477–488, 1970.

[30] D. A. Hsu. Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association*, 74:31–40, 1979.

[31] C. Inclan and G. C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[32] R. Jaxk, J. Chen, X. L. Wang, R. Lund, and L. QiQi. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 6:900–915, 2007.

[33] H. Jeffreys. *The theory of probability.* Oxford, 1961.

[34] R. Killick, I. A. Eckley, K. Ewans, and P. Jonathan. Detection of changes in the characteristics of oceanographic time-series using change point analysis. *(In submission)*, 2009.

[35] D. W. Kwon, K. Ko, M. Vannucci, A. L. N. Reddy, and S. Kim. Wavelet methods for the detection of anomalies and their application to network traffic analysis. *Quality and reliability engieneering international*, 22:953–969, 2006.

[36] M. Lavielle and E. Lebarbier. An application of MCMC methods for the multiple change-points problem. *Signal Processing*, 81:39–53, 2001.

[37] P. Lio and M. Vannucci. Wavelet change-point prediction of transmembrane proteins . *Bioinformatics*, 16(4):376–382, 2000.

[38] J. S. Liu and C. E. Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15:38–52, 1999.

[39] E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

[40] A. N. Pettitt. A non-parametric approach to the change-point problem. *Applied Statistics*, 28:126–135, 1979.

[41] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald. Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing*, 50:747–758, 2002.

[42] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[43] A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.

[44] A. Sen and M. S. Srivastava. On tests for detecting change in mean. *The Annals of Statistics*, 3(1):98–108, 1975.

[45] A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, 8:26–51, 1963.

[46] V. Spokoiny. Multiscale local change point detection with applications to value-at-risk. *The Annals of Statistics*, 37:1405–1436, 2009.

[47] A. G. Tartakovsky and V. V. Veeravalli. General asymptotic Bayesian theory of quickest change detection. *Theory of Probability and Its Applications*, 49:458–497, 2004.

[48] E. S. Venkatraman. *Consistency results in multiple change-point problems.* PhD thesis, Stanford University, 1993.

[49] L. J. Vostrikova. Detecting disorder in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59, 1981.

[50] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[51] G. Yan, Z. Xiao, and S. Eidenbenz. Catching instant messaging worms with change-point detection techniques. In *Proceedings of the USENIX workshop on large-scale exploits and emergent threats*, 2008.

[52] T. Y. Yang and L. Kuo. Bayesian binary segmentation procedure for a Poisson process with multiple changepoints. *Journal of Computational and Graphical Statistics*, 10:772–785, 2001.

[53] Y. Yao. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics*, 12:1434–1447, 1984.

[54] Y. Yao. Estimating the number of change-points via Schwarz's criterion. *Statistics and Probability Letters*, 6:181–189, 1988.