Bayesian Analysis (2017) **12**, Number 1, pp. 193–218

Bayesian Detection of Abnormal Segments in Multiple Time Series

Lawrence Bardwell[*] and Paul Fearnhead[†]

Abstract. We present a novel Bayesian approach to analysing multiple time-series with the aim of detecting abnormal regions. These are regions where the properties of the data change from some normal or baseline behaviour. We allow for the possibility that such changes will only be present in a, potentially small, subset of the time-series. We develop a general model for this problem, and show how it is possible to accurately and efficiently perform Bayesian inference, based upon recursions that enable independent sampling from the posterior distribution. A motivating application for this problem comes from detecting copy number variation (CNVs), using data from multiple individuals. Pooling information across individuals can increase the power of detecting CNVs, but often a specific CNV will only be present in a small subset of the individuals. We evaluate the Bayesian method on both simulated and real CNV data, and give evidence that this approach is more accurate than a recently proposed method for analysing such data.

Keywords: BARD, changepoint detection, copy number variation, PASS.

1 Introduction

In this paper we consider the problem of detecting abnormal (or outlier) segments in multivariate time series. We assume that the series has some normal or baseline behaviour but that in certain intervals or segments of time a subset of the dimensions of the series has some kind of altered or abnormal behaviour. By the term abnormal behaviour we mean some change in distribution of the data away from the baseline distribution. For example, this could include a change in mean, variance or auto-correlation structure. In particular our work is concerned with situations where the size of this subset is only a small proportion of the total number of dimensions. We attempt to do this in a fully Bayesian framework.

This problem is increasingly common across a range of applications where the detection of abnormal segments (sometimes known as recurrent signal segments) is of interest (particularly in high dimensional and/or very noisy data). Some example applications include the analysis of the correlations between sensor data from different vehicles (Spiegel et al., 2011) or for intrusion detection in large interconnected computer networks (Qu et al., 2005). Another related application involves detecting common and potentially more subtle objects in a number of images, for example Jin (2004) and the references therein look at this in relation to multiple images taken of astronomical bodies.

We will focus in particular on one specific example of this type of problem, namely that of detecting copy number variants (CNV's) in DNA sequences. A CNV is a type of

[*] STOR-i CDT, Lancaster University, UK, l.bardwell@lancaster.ac.uk
[†] STOR-i CDT, Lancaster University, UK, p.fearnhead@lancaster.ac.uk

DOI: 10.1214/16-BA998

structural variation that results in a genome having an abnormal (generally $\neq 2$) number of copies of a segment of DNA, such as a gene. Understanding these is important as these variants have been shown to account for much of the variability within a population. For a more detailed overview of this topic see Zhang (2010); Jeng et al. (2013) and the references therein.

Data on CNVs for a given cell or individual is often in the form of "log-R ratios" for a range of probes, each associated with different locations along the genome. These are calculated as log base 2 of the ratio of the measured probe intensity to the reference intensity for a given probe. Normal regions of the genome would have log-R ratios with a mean of 0, whereas CNVs would have log-R ratios with a mean that is away from zero.

Figure 1 gives an example of such data from 6 individuals. We can see that there is substantial noise in the data, and each CNV may cover only a relatively small region of the genome. Both these factors mean that it can be difficult to accurately detect CNVs by analysing data from a single individual or cell. To increase the power to identify CNVs we can pool information by jointly analysing data from multiple individuals. However this is complicated as a CNV may be observed for only a subset of the individuals. For example, for the data in Figure 1, which shows data from a small portion of chromosome 16, we have identified a single CNV which affects only three individuals. This can seen by the raised means (indicated by the red lines) in these three series for a segment of data. By comparison, the other individuals are unaffected in this segment.

Whilst there has been substantial research into methods for detecting outliers (Tsay et al., 2000; Galeano et al., 2006) or abrupt changes in data (Olshen et al., 2004; Jandhyala et al., 2013; Wyse et al., 2011; Frick et al., 2014), the problem of identifying outlier regions in just a subset of dimensions has received less attention. Exceptions include methods described in Zhang et al. (2010) and Siegmund et al. (2011). However Jeng et al. (2013) argue that these methods are only able to detect common variants, that is abnormal segments for which a large proportion of the dimensions have undergone the change. Jeng et al. (2013) propose a method, the PASS algorithm, which is also able to detect rare variants.

The methods of Siegmund et al. (2011) and Jeng et al. (2013) are based on defining an appropriate test-statistic for whether a region is abnormal for a subset of dimensions, and then recursively using this test-statistic to identify abnormal regions. As such the output of these methods is a list of estimated abnormal regions. Here we introduce a Bayesian approach to detecting abnormal regions. This is able to not only give estimates of the number and location of the abnormal regions, but to also give measures of uncertainty about these. We show how it is possible to efficiently simulate from the posterior distribution of the number and location of abnormal regions, through using recursions similar to those from multiple changepoint detection (Barry and Hartigan, 1992; Fearnhead, 2006; Fearnhead and Vasileiou, 2009). We call the resulting algorithm, Bayesian Abnormal Region Detector (BARD).

The outline of the paper is as follows. In the next section we introduce our model, both for the general problem of detecting abnormal regions, and also for the specific

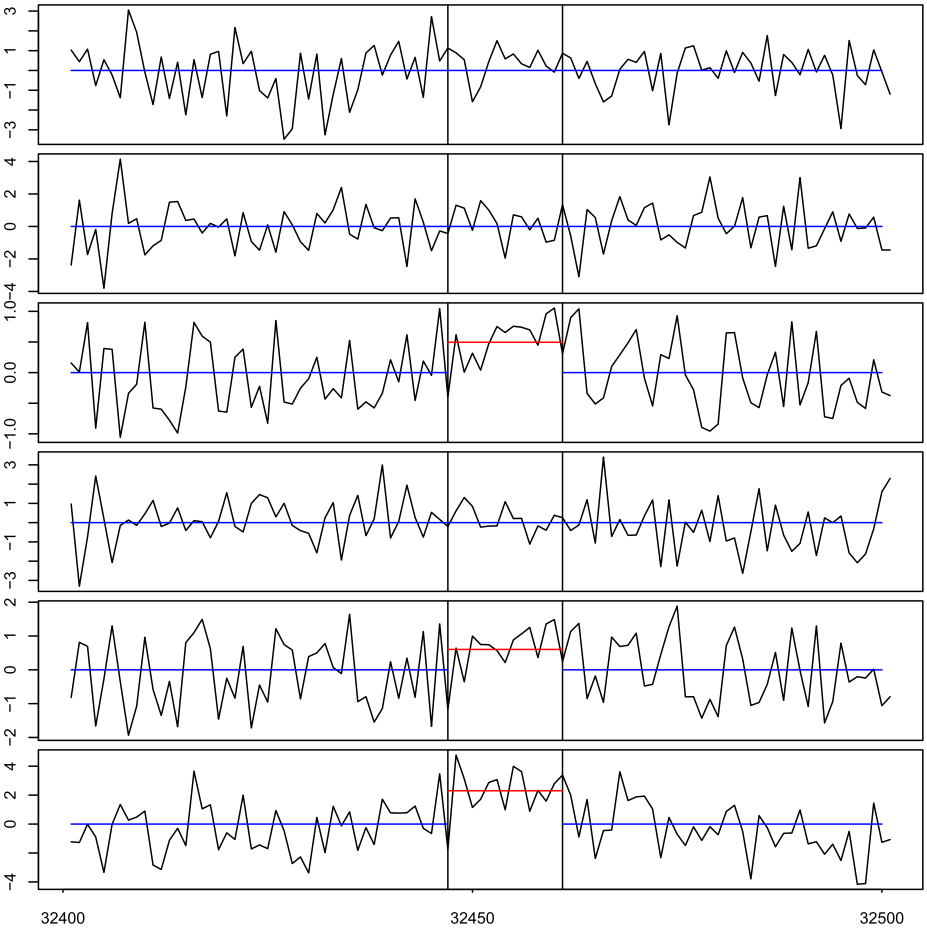


Figure 1: Log-R ratios from 6 individuals for a small portion of chromosome 16. We indicate the baseline level (mean zero) by a horizontal line in blue and the identified CNV (abnormal region) is highlighted between two vertical black lines with the mean of the affected individuals in red.

CNV application. In Section 3 we derive the recursions that enable us to draw iid samples from the posterior, as well as a simple approximation to these recursions that results in an algorithm, BARD, that scales linearly with the length of data set. We then present theoretical results that show that BARD can consistently estimate the absence of abnormal segments, and the location of any abnormal segments, and is robust to some mis-specification of the priors. In Section 5 we evaluate BARD for the CNV application on both simulated and real data. Our results suggest that BARD is more accurate than PASS, particularly in terms of having fewer false positives. Furthermore, we see evidence that posterior probabilities are well-calibrated and hence are accurately representing the uncertainty in the inferences. The paper ends with a discussion.

2 The Model

We shall now describe the details of our model. Consider a multiple time series of dimension d and length n, $\mathbf{Y}_{1:n} = (\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n)$ where $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,d})^T$. We model this data through introducing a hidden state process, $X_{1:n}$. The hidden state process will contain information about where the abnormal segments of the data are. Our model is defined through specifying the distribution of the hidden state process, $p(x_{1:n})$, and the conditional distribution of the data given the state process, $p(\mathbf{y}_{1:n}|x_{1:n})$. These are defined in Sections 2.1 and 2.2 respectively.

Our interest lies in inference about this hidden state process given the observations. This involves calculating the posterior distribution for the states

$$p(x_{1:n}|\mathbf{y}_{1:n}) \propto p(x_{1:n}, \mathbf{y}_{1:n}) = p(x_{1:n})p(\mathbf{y}_{1:n}|x_{1:n}). \qquad (1)$$

It should be noted that these probabilities will depend on a set of hyper-parameters. These parameters are initially assumed to be known, however we will later discuss performing inference for them.

2.1 Hidden State Model

The hidden state process will define the location of the abnormal segments. We will model the location of these segments through a renewal process. The length of a given segment is drawn from some distribution which depends on the segment type, and is independent of all other segment lengths. We assume a normal segment is always followed by an abnormal segment, but allow for either a normal or abnormal segment to follow an abnormal one. The latter is because each abnormal segment may be abnormal in a different way, for example with different subsets of the time-series being affected. This will become clearer when we discuss the likelihood model in Section 2.2.

To define such a model we need distributions for the lengths of normal and abnormal segments. We denote the cumulative distribution functions of these lengths by $G_N(t)$ and $G_A(t)$ respectively. We also need to specify the probability that an abnormal segment is followed by either a normal or abnormal segment. We denote these probabilities as π_N and π_A respectively, with $\pi_N = 1 - \pi_A$.

Note that the first segment for the data will have a different distribution to other segments as it may have started at some time prior to when we started collecting data. We can define this distribution in a way that is consistent with our underlying model by assuming the process for the segments is at stationarity and that we start observing it at an arbitrary time. Renewal theory (Cox, 1962) then gives the distribution function for the length of the first segment. If the first segment is normal, then we define its cumulative distribution function as

$$G_{0N}(t) = \sum_{s=1}^{t} \frac{1 - G_N(s)}{E_N},$$

where E_N is the expected length of a normal segment. The cumulative distribution function for the first segment conditional on it being abnormal, $G_{0A}(t)$, is similarly defined.

Formally, we define our hidden state process X_t as $X_t = (C_t, B_t)$ where C_t is the end of the previous segment prior to time t and B_t is the type of the current segment. So $C_t \in \{0, \dots, t-1\}$ with $C_t = 0$ denoting that the current segment is the first segment. We use the notation that $B_t = N$ if the current segment is normal, and $B_t = A$ if not. This state process is Markov, and thus we can write

$$\begin{aligned} p(x_{1:n}) &= p(c_{1:n}, b_{1:n}) \\ &= \Pr(C_1 = c_1, B_1 = b_1) \prod_{i=1}^{n-1} \Pr(C_{i+1} = c_{i+1}, B_{i+1} = b_{i+1} | C_i = c_i, B_i = b_i). \end{aligned} \tag{2}$$

The decomposition in (2) gives us two aspects of the process to define, namely the transition probabilities $\Pr(C_{i+1} = c_{i+1}, B_{i+1} = b_{i+1} | c_i, b_i)$ and the initial distribution, $\Pr(C_1 = c_1, B_1 = b_1)$.

Firstly consider the transition probabilities. Now either $C_{t+1} = C_t$ or $C_{t+1} = t$ depending on whether a new segment starts between time t and $t+1$. The probability of a new segment starting is just the conditional probability of a segment being of length $t - C_t$ given that is at least $t - C_t$. If $C_{t+1} = C_t$, then we must have $B_{t+1} = B_t$, otherwise the distribution of the type of the new segment depends on the type of the previous segment as described above.

Thus for $i = 1, \dots, t-1$ we have

$$\begin{aligned} \Pr(C_{t+1} = j, B_{t+1} = k | C_t = i, B_t = N) &= \begin{cases} \frac{1-G_N(t-i)}{1-G_N(t-i-1)} & \text{if } j = i \text{ and } k = N, \\ \frac{G_N(t-i)-G_N(t-i-1)}{1-G_N(t-i-1)} & \text{if } j = t \text{ and } k = A, \\ 0 & \text{otherwise,} \end{cases} \\ \Pr(C_{t+1} = j, B_{t+1} = k | C_t = i, B_t = A) &= \begin{cases} \frac{1-G_A(t-i)}{1-G_A(t-i-1)} & \text{if } j = i \text{ and } k = A, \\ \pi_A \left(\frac{G_A(t-i)-G_A(t-i-1)}{1-G_A(t-i-1)} \right) & \text{if } j = t \text{ and } k = A, \\ \pi_N \left(\frac{G_A(t-i)-G_A(t-i-1)}{1-G_A(t-i-1)} \right) & \text{if } j = t \text{ and } k = N, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3}$$

For $i = 0$, that is when $C_t = 0$, we replace $G_N(\cdot)$ and $G_A\cdot$ with $G_{0N}(\cdot)$ and $G_{0A}(\cdot)$ respectively.

Finally we need to define the initial distribution for $X_1 = (B_1, C_1)$. Firstly note that $C_1 = 0$ so we need only the distribution of B_1. We define this as the stationary distribution of the B_t process. This is (see for example Theorem 5.6 of Kulkarni, 2012)

$$\Pr(B_1 = N) = \frac{\pi_N E_N}{\pi_N E_N + E_A}, \quad \Pr(B_1 = A) = 1 - \Pr(B_1 = N),$$

where E_N and E_A are the expected lengths of normal and abnormal segments respectively.

2.2 Likelihood Model

The hidden process $X_{1:n}$ described above partitions the time interval into contiguous non-overlapping segments each of which is either normal, N, or abnormal, A. Now conditional on this process we want to define a likelihood for the observations, $p(\mathbf{y}_{1:n}|x_{1:n})$.

For many applications it is natural to assume a conditional independence property between segments: this means that if we knew the locations of segments and their types then data from different segments are independent. This assumption is key to the algorithms we later introduce to sample from the posterior. Thus when we condition on C_t and B_t the likelihood for the first t observations factorises as follows

$$p(\mathbf{y}_{1:t}|C_t = j, B_t) = p(\mathbf{y}_{1:j}|C_t = j, B_t)p(\mathbf{y}_{j+1:t}|C_t = j, B_t). \tag{4}$$

The second term in equation (4) is the marginal likelihood of the data, $\mathbf{Y}_{j+1:t}$, given it comes from a segment that has type B_t. We introduce the following notation for these segment marginal likelihoods, where for $s \geq t$,

$$\begin{aligned} P_N(t, s) &= \Pr(\mathbf{y}_{t:s}|C_s = t-1, B_s = N), \\ P_A(t, s) &= \Pr(\mathbf{y}_{t:s}|C_s = t-1, B_s = A), \end{aligned} \tag{5}$$

and define $P_N(t, s) = 1$ and $P_A(t, s) = 1$ if $s < t$.

Now using the above factorisation we can write down the likelihood conditional on the hidden process. Note that we can condition on X_t rather than the full history $X_{1:n}$ in each of the factors in (6) due to the conditional independence assumption on the segments

$$\begin{aligned} p(\mathbf{y}_{1:n}|x_{1:n}) &= \prod_{t=1}^{n} p(\mathbf{y}_t|x_{1:n}, \mathbf{y}_{1:(t-1)}) \\ &= \prod_{t=1}^{n} p(\mathbf{y}_t|C_t, B_t, \mathbf{y}_{(C_t+1):(t-1)}). \end{aligned} \tag{6}$$

The terms on the right-hand side of equation (6) can then be written in terms of the segment marginal likelihoods

$$p(\mathbf{y}_t|C_t, B_t, \mathbf{y}_{(C_t+1):(t-1)}) = \frac{P_{B_t}(C_t + 1, t)}{P_{B_t}(C_t + 1, t-1)}. \tag{7}$$

Thus our likelihood is specified through defining appropriate forms for the marginal likelihoods for normal and abnormal segments.

Model for Data in Normal Segments

For a normal segment we model that the data for all dimensions of the series are realisations from some known distribution, $\mathcal{D}$, and these realisations are independent over both time and dimension. Denote the density function of the distribution $\mathcal{D}$ as $f_{\mathcal{D}}(\cdot)$. We can write down the segment marginal likelihood as

$$P_N(t,s) = \prod_{k=1}^{d} \prod_{i=t}^{s} f_{\mathcal{D}}(y_{i,k}). \tag{8}$$

Model for Data in Abnormal Segments

For abnormal segments our model is that data for a subset of the dimensions are drawn from $\mathcal{D}$, with the data for the remaining dimensions being independent realisations from a different distribution, $\mathcal{P}_\theta$, which depends on a segment specific parameter θ. We denote the density function for this distribution as $f_{\mathcal{P}}(\cdot|\theta)$.

Our model for which dimensions have data drawn from $\mathcal{P}_\theta$ is that this occurs for dimension k with probability p_k, independently of the other dimensions. Thus if we have an abnormal segment with data $\mathcal{Y}_{t:s}$, with segment parameter θ, the likelihood of the data associated with the kth dimension is

$$p_k \prod_{i=t}^{s} f_{\mathcal{P}}(y_{i,k}|\theta) + (1-p_k) \prod_{i=t}^{s} f_{\mathcal{D}}(y_{i,k}).$$

Thus by independence over dimension

$$p(\mathbf{y}_{t:s}|\theta) = \prod_{k=1}^{d} \left(p_k \prod_{i=t}^{s} f_{\mathcal{P}}(y_{i,k}|\theta) + (1-p_k) \prod_{i=t}^{s} f_{\mathcal{D}}(y_{i,k}) \right).$$

Our model is completed by a prior for θ, $\pi(\theta)$. To find the marginal likelihood $P_A(t,s)$ we need to integrate out θ from $p(\mathbf{y}_{t:s}|\theta)$

$$P_A(t,s) = \int p(\mathbf{y}_{t:s}|\theta)\pi(\theta)\,\mathrm{d}\theta. \tag{9}$$

In practice this integral will need to be calculated numerically, which is feasible if θ is low-dimensional.

CNV Example

In Section 1 we discussed the copy number variant (CNV) application and showed some real data in Figure 1. From the framework described above we now need to specify a model for normal and abnormal segments. Following Jeng et al. (2013) we model the data as being normally distributed with constant variance but differing means either zero or μ depending on whether we are in a normal or abnormal segment. This model also underpins the simulation studies that we present in Section 5.

Using the notation from the more general framework discussed above the two distributions for normal and abnormal segments are

$$\mathcal{D} \sim N(0,\sigma^2)$$

$$\mathcal{P}_\mu \sim N(\mu, \sigma^2).$$

We assume that the variance σ^2 is constant and known. In practice we estimate this quantity using the robust median absolute deviation estimator as recommended in Jeng et al. (2013).

Having specified these two distributions we then need to calculate marginal likelihoods for normal and abnormal segments given by equations (8) and (9) respectively. Calculating the marginal likelihood for a normal segment is simple because of independence over time and dimension as shown in equation (8). However calculating $P_A(\cdot,\cdot)$ is more challenging, as there is no conjugacy between $p(\mathbf{y}|\mu)$ and $\pi(\mu)$ so we can only numerically approximate the integral. Calculating the numerical approximation is fast as it is a one-dimensional integral.

In the simulation studies and results we take the prior for μ to be uniform on a region that excludes values of μ close to zero. For CNV data such a prior seems reasonable empirically (see Figure 2c) and also because we expect CNV's to correspond to a change in mean level of at least $\log(3/2)$ and can be both positive or negative.

3 Inference

We now consider performing inference for the model described in Section 2. Firstly a set of recursions to perform this task exactly are introduced and then an approximation is considered to make this procedure computationally more efficient.

3.1 Exact On-Line Inference

We follow the method of Fearnhead and Vasileiou (2009) in developing a set of recursions for the posterior distribution of the hidden state, the location of the start of the current segment and its type, at time t given that we have observed data upto time t, $p(x_t|\mathbf{y}_{1:t}) = p(c_t, b_t|\mathbf{y}_{1:t})$, for $t \in \{1, 2, \ldots, n\}$. These are known as the filtering distributions. Eventually we will be able to use these to simulate from the full posterior, $p(x_{1:n}|\mathbf{y}_{1:n})$.

To find these filtering distribution we develop a set of recursions that enable us to calculate $p(c_{t+1}, b_{t+1}|\mathbf{y}_{(1:t+1)})$ in terms of $p(c_t, b_t|\mathbf{y}_{1:t})$. These recursions are analogous to the forward-backward equations widely used in analysing Hidden Markov models.

There are two forms of these recursions depending on whether $C_{t+1} = j$ for $j < t$ or $C_{t+1} = t$. We derive the two forms separately. Consider the first case. For $j < t$ and $k \in \{N, A\}$,

$$\begin{aligned}
&p(C_{t+1} = j, B_{t+1} = k|\mathbf{y}_{1:(t+1)}) \\
&\quad \propto p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}, C_{t+1} = j, B_{t+1} = k)p(C_{t+1} = j, B_{t+1} = k|\mathbf{y}_{1:t}) \\
&\quad = \left(\frac{P_k(j+1, t+1)}{P_k(j+1, t)}\right) \Pr(C_{t+1} = j, B_{t+1} = k|C_t = j, B_t = k)p(C_t = j, B_t = k|\mathbf{y}_{1:t}),
\end{aligned}$$

where the first term in the last expression is the conditional likelihood from equation (7). The second two terms use the fact that there has not been a new segment and hence $C_{t+1} = C_t$ and $B_{t+1} = B_t$.

Now for the second case, when $C_{t+1} = t$,

$$\begin{aligned} &p(C_{t+1} = t, B_{t+1} = k|\mathbf{y}_{1:t}) \\ &\qquad = \sum_{i=0}^{t-1} \sum_{l \in \{N,A\}} p(C_t = i, B_t = l|\mathbf{y}_{1:t}) \Pr(C_{t+1} = t, B_{t+1} = k|C_t = i, B_t = l). \end{aligned}$$

Thus, as $p(\mathbf{y}_{t+1}|C_{t+1} = t, B_{t+1} = k, \mathbf{y}_{1:t}) = P_k(t+1, t+1)$, the filtering recursion is;

$$\begin{aligned} &p(C_{t+1} = t, B_{t+1} = k|\mathbf{y}_{1:(t+1)}) \\ &\qquad \propto P_k(t+1, t+1) \sum_{i=0}^{t-1} \sum_{l \in \{N,A\}} p(C_t = i, B_t = l|\mathbf{y}_{1:t}) \\ &\qquad\quad \times \Pr(C_{t+1} = t, B_{t+1} = k|C_t = i, B_t = l). \end{aligned}$$

These recursions are initialised by $p(C_1 = 0, B_1 = k|\mathbf{y}_1) \propto \Pr(B_1 = k) P_k(1, 1)$ for $k \in \{N, A\}$.

3.2 Approximate Inference

The support of the filtering distribution $p(c_t, b_t|\mathbf{y}_{1:t})$ has $2t$ points. Hence, calculating $p(c_t, b_t|\mathbf{y}_{1:t})$ exactly is of order t both in terms of computational and storage costs. The cost of calculating and storing the full set of filtering distributions $t = 1, 2, \ldots, n$ is thus of order n^2. For larger data sets this exact calculation can be prohibitive. A natural way to make this more efficient is to approximate each of the filtering distributions by distributions with a fewer number of support points. In practice such an approximation is feasible as many of the support points of each filtering distribution have negligible probability. If we removed these points then we could greatly increase the speed of our algorithm without sacrificing too much accuracy.

We use the stratified rejection control (SRC) algorithm (Fearnhead and Liu, 2007) to produce an approximation to the filtering distribution with potentially fewer support points at each time-point. This algorithm requires the choice of a threshold, $\alpha \geq 0$. At each iteration the SRC algorithm keeps all support points which have a probability greater than α. For the remaining particles the probability of them being removed is proportional to their associated probability and the resampling is done in a stratified manner. This algorithm has good theoretical properties in terms of the error introduced at each resampling step, measured by the Kolmogorov Smirnov distance, being bounded by α.

3.3 Simulation

Having calculated and stored the filtering distributions, either exactly or approximately, simulating from the posterior is straightforward. This is performed by simulating the hidden process backwards in time (Carter and Kohn, 1994). First we simulate $X_n = (C_n, B_n)$ from the final filtering distribution $p(c_n, b_n|\mathbf{y}_{1:n})$. Assume we simulate $C_n = t$. Then, by definition of the hidden process, we have $C_s = t$ and $B_s = B_n$ for $s = t+1, \ldots, n-1$, as these time-points are all part of the same segment. Thus we next need to simulate C_t, from its conditional distribution given C_{t+1}, B_{t+1} and $\mathbf{Y}_{1:n}$,

$$\begin{aligned}
&p(c_t, b_t|C_{t+1} = t, B_{t+1}, \mathbf{y}_{1:n})\\
&\qquad \propto p(c_t, b_t, C_{t+1} = t, B_{t+1}, \mathbf{y}_{1:n})\\
&\qquad = p(c_t, b_t)\,\Pr(C_{t+1} = t, B_{t+1}|C_t, B_t) p(\mathbf{y}_{1:n}|C_t, B_t, C_{t+1} = t, B_{t+1})\\
&\qquad \propto p(c_t, b_t)\,\Pr(C_{t+1} = t, B_{t+1}|C_t, B_t) p(\mathbf{y}_{1:t}|C_t, B_t)\\
&\qquad \propto p(c_t, b_t|\mathbf{y}_{1:t})\,\Pr(C_{t+1} = t, B_{t+1}|C_t, B_t).
\end{aligned}$$

We then repeat this process, going backwards in time until we simulate $C_t = 0$. From the simulated values we can extract the location and type of each segment.

3.4 Hyper-Parameters

As mentioned earlier in Section 2 the posterior of interest (1) depends upon a vector of hyper-parameters which we now label as Ψ. In Section 5, Ψ contains the parameters for the LOS distributions for the two differing types of segments which determine the cdf's $G_N(\cdot)$ and $G_A(\cdot)$. However we could extend Ψ to account for the hyperparameters for the prior on μ or, if we did not assume a common and known variance for the data, the variance for each time-series.

We use two approaches to estimating these hyper-parameters. The first is to maximise the marginal-likelihood for the hyper-parameters, which we can do using Monte Carlo EM (MCEM). For general details on MCEM see Levine and Casella (2001). Although convergence of the hyper-parameters is quite rapid in the examples we look at in Section 5, for very large data sets a cruder but faster alternative is to initially segment the data using a different method to ours and then use information from this segmentation to inform the choice of hyper-parameter values. The alternative method we use is the PASS method of Jeng et al. (2013) and discussed in detail in Section 5.

3.5 Estimating a Segmentation

We have described how to calculate the posterior density $p(x_{1:n}|\mathbf{y}_{1:n})$ from which we can easily draw a large number of samples. However we often want to report a single estimated "best" segmentation of the data. We can define such a segmentation using Bayesian decision theory (Berger, 1985). This involves defining a loss function which determines the cost of us making a mistake in our estimate of the true quantity which we then seek to minimise. There are various choices of loss function we could use (see Yau and Holmes, 2010), but we use a loss that is a sum of a loss for estimating whether each location is abnormal or not. If $L(\tilde{b}_t|b_t)$ gives the cost of making the decision that the state at time t is $\tilde{b}_t$ when in fact it is b_t, then:

$$L(\tilde{b}_t|b_t) = \begin{cases} 1 & \text{if } \tilde{b}_t = \text{A and } b_t = \text{N} \\ \gamma & \text{if } \tilde{b}_t = \text{N and } b_t = \text{A} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The inclusion of γ allows us to vary the relative penalty for false positives as compared to false negatives. Under this loss we estimate $\hat{b}_t = N$ if $\pi(b_t = A) < 1/(1+\gamma)$ or $\hat{b}_t = A$ otherwise.

4 Asymptotic Consistency

We will now consider the asymptotic properties of the method as d, the number of time-series, increases. Our aim is to study the robustness of inferences to the choice of prior for the abnormal segments, and the estimate of p_d, allowing for abnormal segments that are rare. We will assume that each time-series is of fixed length n. Following Jeng et al. (2013), to consider the influence of rare abnormal segments, we will let the proportion of sequences that are abnormal in an abnormal segment to decrease as d increases.

Our assumptions on how the data is generated is that there are a fixed number and location of abnormal segments. We will assume the model of Section 2.2 with, without loss of generality, $\sigma^2 = 1$ (for $\sigma^2 \neq 1$ we can just normalise the data). So if $B_t = N$, then $Y_{i,j} \sim N(0,1)$. If $(t,\ldots,s)$ is an abnormal segment then it has an associated mean, $\mu_0 \neq 0$. For each $j = 1,\ldots,d$, independently with probability α_d, $Y_{i,j} \sim N(\mu,1)$ for $i = t,\ldots,s$; otherwise $Y_{i,j} \sim N(0,1)$ for $i = t,\ldots,s$.

We fit the model of Section 2, assuming the correct likelihood for data in normal and abnormal segments. For each abnormal segment we will have an independent prior for the associated mean, $\pi(\mu)$. Our assumptions on $\pi(\mu)$ is that its support is a subset of $\{[-b,-a],[a,b]\}$ for some $a > 0$ and $b < \infty$, and it places non-zero probability on both positive and negative values of μ. The model we fit will assume a specified probability, p_d, of each sequences being abnormal within each abnormal segment. Note that we do not require $p_d = \alpha_d$, the true probability, but we do allow the choice of this parameter to depend on d.

The lemmas used in the proof of the following two theorems can be found in the Supplementary material (see Bardwell and Fearnhead, 2016).

Theorem 1. *Assume the model for the data and the constraints on the prior specified above. Let $\mathcal{E}$ be the event that there are no abnormal segments, and $\mathcal{E}^c$ its complement. If there are no abnormal segments and $d \to \infty$, with $1/p_d = O(d^{\frac{1}{2}-\epsilon})$ for some $\epsilon > 0$, then*

$$\Pr(\mathcal{E}^c|\mathbf{y}_{1:n}) \to 0,$$

in probability.

Proof. As n is fixed, we have a fixed number of possible segmentations. We will show that the posterior probability of each possible segmentation with at least one abnormal segment is $o_p(1)$ as $d \to \infty$.

For time-series k let $P_{N,k}(t,s)$ denote the likelihood of the data $y_{t,k},\dots,y_{s,k}$ assuming this is a normal segment; and let $P_{A,k}(t,s;\mu)$ be the marginal likelihood of the same data given that it is drawn from independent Gaussian distributions with mean μ. Then if we have a segmentation with m abnormal segments, with the ith abnormal segment from t_i to s_i, the ratio of the posterior probability of this segmentation to the posterior probability of $\mathcal{E}$ is

$$K\prod_{i=1}^{m}\left(\int\left\{\prod_{k=1}^{d}\frac{P_{A,k}(t_m,s_m;\mu)}{P_{N,k}(t_m,s_m)}\right\}\pi(\mu)\mathrm{d}\mu\right),$$

where K is the ratio of the prior probabilities of these two segmentations. So it is sufficient to show that for all $t\le s$,

$$\int\left\{\prod_{k=1}^{d}\frac{P_{A,k}(t,s;\mu)}{P_{N,k}(t,s)}\right\}\pi(\mu)\mathrm{d}\mu\to 0 \tag{11}$$

in probability as $d\to\infty$.

Our limit involves treating the data as random. Each term in this product is then random, and of the form

$$\frac{P_{A,k}(t,s;\mu)}{P_{N,k}(t,s)}=1+p_d\left(\exp\left\{\mu\sum_{u=t}^{s}\left(Y_{k,u}-\frac{\mu}{2}\right)\right\}-1\right). \tag{12}$$

By applying Lemma 1.4 separately to positive and negative values of μ, we have that this tends to 0 with probability 1 as $d\to\infty$. This is true for all possible segmentations with at least one abnormal segments. As n is fixed, there are a finite number of such segments, so the result follows. □

Theorem 2 tells us that the posterior probability of misclassifying a time point as normal when it is abnormal tends to zero as more time-series are observed.

Theorem 2. *Assume the model for the data and the constraints on the prior specified above. Fix any position t, and consider the limit as $d\to\infty$, with $dp_d^2\to\infty$ and either*

(i) $p_d=o(\alpha_d)$; or

(ii) if μ_0 is the mean associated with the abnormal sequences at position t, then there exists a region A such that the prior probability associated with $\mu\in A$ is non-zero, and for all $\mu\in A$ and for sufficiently large d

$$\alpha_d\left(e^{\mu\mu_0}-1\right)-\frac{p_d}{2}\left(e^{\mu^2}-1\right)>0.$$

Then if $B_t=A$

$$\Pr(B_t=N|\mathbf{y}_{1:n})\to 0.$$

in probability.

Proof. We will show that each segmentation with $B_t = N$ has posterior probability that tends to 0 in probability as $d \to \infty$. For each segmentation with $B_t = N$ we will compare its posterior probability with one which is identical except for the addition of an abnormal segmentant, of length 1, at location t. The ratio of posterior probabilities of these two segmentations will be

$$K \left(\int \left\{ \prod_{k=1}^{d} \frac{P_{A,k}(t,t;\mu)}{P_{N,k}(t,t)} \right\} \pi(\mu) \mathrm{d}\mu \right),$$

where K is a constant that depends on the prior for the segmentations. We require that this ratio tends to infinity in probability as $d \to \infty$. Under both conditions (i) and (ii) above this follows immediately from Lemma 2.2. For case (i) we are using the fact that the prior places positive probability both on μ being positive and negative, and for μ the same sign as μ_0 we have that $e^{\mu\mu_0} > 1$. □

This result shows some robustness of the Bayesian approach to the choice of prior. Consider a prior on the mean for an abnormal segment that has strictly positive density for values in $\{[-b,-a],[a,b]\}$ for $a > 0$. Then for any true mean, μ_0 with $|\mu_0| \geq a$, we will consistently estimate the segment as abnormal provided the assumed or estimated probability of a sequence being abnormal is less than twice the true value. Thus we want to choose a to be the smallest absolute value of the mean of an abnormal segment we expect or wish to detect. The choice of b is less important, in that it does not affect the asymptotic consistency implied by the above theorem.

Furthermore we do not need to specify p_d exactly for consistency – the key is not to over-estimate the true proportion of abnormal segments by more than a factor of two. We could set $p_d = Kd^{-1/2+\epsilon}$ for some constants $K, \epsilon > 0$ and ensure that asymptotically we will consistently estimate the absence of abnormal segments (Theorem 1) and the location of any abnormal segments (Theorem 2) the true proportion of abnormal segments decays at a rate that is slower than $d^{-1/2+\epsilon}$.

5 Results

We call the method introduced in Sections 2 and 3 BARD: Bayesian Abnormal Region Detector. We now evaluate BARD on both simulated and real CNV data. Our aim is to both investigate its robustness to different types of model mis-specification, and to compare its performance with a recently proposed method for analysing such CNV data.

The simulation studies we present are based on the concrete example in Section 2.2, namely the change in mean model for Normally distributed data. For inference we assume that the LOS distributions, S_N and S_A, to be Negative binomial and the prior probability of a particular dimension k being abnormal p_k as the same for all $k = \{1,2,\dots,d\}$. For all the simulation studies we present we used MCEM on a single replicate of the simulated data set to get estimates for the hyper-parameters for the LOS distribution, but fixed p_k. Data for normal segments are IID standard Gaussian, and for abnormal segments data from dimensions that are abnormal are Gaussian with variance 1 but mean μ drawn from some prior $\pi(\mu)$. Below we consider the effect of varying the choice of prior used for simulating the data and that assumed within BARD. In implementing BARD we used the SRC method of resampling described in Section 3.2 with a value of $\alpha = 10^{-4}$, we found this value of α gave a good trade off between accuracy and computational cost.

To get an explicit segmentation from BARD we use the asymmetric loss function (10) with a value of $\gamma = 1/3$.

As a benchmark for comparison we also analyse all data sets using the Proportion Adaptive Segment Selection procedure (PASS) from Jeng et al. (2013). This was implemented using an R package called `PASS` which we obtained from the authors website. At its most basic level the PASS method involves evaluating a test statistic for different segments of the data. After these evaluations the values of the statistic that exceed a certain pre-specified threshold are said to be significant and the segments that correspond to these values are the identified abnormal segments. This threshold is typically found by simulating data sets with no abnormal segments and then choosing the threshold which gives a desired type 1 error, here we take this error to be 0.05 in the simulation studies. The PASS algorithm considers all segments that are shorter than a pre-defined length. To avoid excessive computational costs this length should be as small as possible, but at least as large as the longest abnormal segment we wish to detect (or believe exists in the data). We ran PASS with this length set to ten-times the largest abnormal segment.

We found that a run of PASS was about twice as fast as one run of BARD. In order to estimate the hyper-parameters using MCEM took between 5 and 20 runs of BARD.

Evaluating a Segmentation

To form a comparison between the two methods we must have some way of evaluating the quality of a particular segmentation with respect to the ground truth. We consider the three most important criteria to be the number of true and false positives and the accuracy in detecting the true positives.

We define a segment to be correctly identified or a true positive if it intersects with the true segment. With this definition in mind then finding the true/false positives is simple. Note that results for false positives are the number of false positive segments per data set. To define the accuracy of an estimated segment compared to the truth it is most intuitive to measure the amount of "overlap" of the segments, this is captured by the dissimilarity measure D_k (13) defined in Jeng et al. (2013).

Let $\hat{\mathbb{I}}$ be the collection of estimated intervals, the accuracy of estimating the kth true segment I_k is given by D_k

$$D_k = \min_{\hat{I}_j \in \hat{\mathbb{I}}} \left\{ 1 - \frac{|\hat{I}_j \cap I_k|}{\sqrt{|\hat{I}_j||I_k|}} \right\} \tag{13}$$

$D_k \in [0,1]$, if $D_k = 0$ then an estimated interval overlaps exactly with segment I_k however if $D_k = 1$ then no estimated intervals overlap with the kth segment, i.e. it hasn't been detected. Smaller values of D indicate a greater overlap.

For all measures we present the average value across the simulated data sets, together with a 95% confidence interval for this average calculated via the bootstrap.

5.1 Simulated Data from the Model

Firstly we analysed data simulated from the model assumed by BARD. A soft maximum on the length of the simulated data of $n = 1000$ was imposed and the number of dimensions fixed at $d = 200$. The LOS distributions were

$$S_N \sim \text{NBinom}(10, 0.1) \text{ and } S_A \sim \text{NBinom}(15, 0.3).$$

Two different distributions were used to generate the altered means for the affected dimensions and we also varied π_N (see Table 1), and for each scenario we implemented the Bayesian method with the correct prior for the abnormal mean, and the correct choice of π_N. The number of affected dimensions for each abnormal segment was fixed at 4% and we fixed p_k to this value. For each scenario we considered we generated 200 data sets.

μ	π_N	Method	Proportion detected	Accuracy	Number of False positives
$U(0.3, 0.7)$	0.5	PASS	0.68 (0.66,0.70)	0.12 (0.11,0.13)	0.80 (0.68,0.93)
		BARD	0.88 (0.87,0.89)	0.077 (0.071,0.084)	0.08 (0.04,0.12)
	0.8	PASS	0.67 (0.65,0.69)	0.13 (0.12,0.14)	1.13 (0.98,1.29)
		BARD	0.78 (0.77,0.80)	0.094 (0.088,0.10)	0.07 (0.04,0.11)
$U(0.5, 0.9)$	0.5	PASS	0.92 (0.91,0.93)	0.074 (0.070,0.078)	1.08 (0.94,1.22)
		BARD	0.98 (0.98,0.99)	0.039 (0.036,0.042)	0.03 (0.01,0.06)
	0.8	PASS	0.94 (0.93,0.95)	0.073 (0.069,0.076)	1.02 (0.88,1.17)
		BARD	0.96 (0.95,0.97)	0.042 (0.040,0.045)	0.02 (0.00,0.04)

Table 1: Scenarios differed in the prior for μ and the value of π_N used to simulate the data. In BARD these same priors were used for the analysis of the data. The results for each scenario are averages across 200 simulated data sets together with 95% confidence interval in brackets.

Results summarising the accuracy of the segmentations obtained by the two methods are shown in Table 1. BARD performed substantially better than PASS here especially with regards to the number of false positives each method found, though this is in part because all the modelling assumptions within BARD are correct for these simulated data sets. It is worth noting that both methods do much better when $\mu \sim U(0.5, 0.9)$ due to the stronger signal present.

We next investigated how robust the results were to our choice for p_k. We just consider $\mu \sim U(0.3, 0.7)$ and $\pi_N = 0.8$ and we vary our choice of p_k from 0.5% to 10%. These results are in Table 2. Whilst, as expected, if we take p_k to be the true value for the data we get the best segmentation, the results are clearly robust to mis-specification

p_k	Proportion detected	Accuracy	Number of False positives
$\frac{1}{200}$	0.65 (0.63,0.67)	0.093 (0.086,0.10)	0.03 (0.005,0.06)
$\frac{4}{200}$	0.76 (0.74,0.78)	0.092 (0.086,0.10)	0.09 (0.05,0.12)
$\frac{8}{200}$	0.77 (0.75,0.78)	0.086 (0.081,0.093)	0.06 (0.03,0.09)
$\frac{12}{200}$	0.76 (0.74,0.78)	0.089 (0.083,0.096)	0.06 (0.03,0.095)
$\frac{16}{200}$	0.74 (0.72,0.76)	0.091 (0.084,0.098)	0.06 (0.03,0.09)
$\frac{20}{200}$	0.72 (0.70,0.74)	0.095 (0.088,0.102)	0.05 (0.02,0.08)

Table 2: The robustness of BARD under a misspecification of p_k taking the prior as $\mu \sim U(0.3, 0.7)$ and $\pi_N = 0.8$ with the true value of p_k being 4%. Values of p_k were varied between 0.5% and 10% and we simulated 200 data sets for each p_k. The results for each scenario are averages across 200 simulated data sets together with 95% confidence interval in brackets.

of p_k. In all cases we still achieve much higher accuracy and fewer false positives than PASS. Apart from the choice $p_k = 1/200$ we also have a higher proportion of correctly detected CNVs than PASS.

We also investigated the robustness to mis-specification of the model for the LOS distribution, and for the distribution of the mean of the abnormal segments. We fixed the position of five abnormal segments at the following time points 200, 300, 500, 600 and 750. Additionally the segments at 200 and 750 were followed by another abnormal segment. Thus we have seven abnormal segments in total. The true LOS distribution for the abnormal segments are in fact Poisson with intensity randomly chosen from the set $\{20, 25, 30, 35, 40\}$. For these abnormal segments the mean value that affected the dimensions was drawn from a Normal distribution with differing means and a fixed variance shown in Table 3. The number of affected dimensions for each of the abnormal segments was also varied randomly from 3–6% of the total number of dimensions ($d =$ 200). For inference, we fixed p_k to 4% for all k and we set the prior for the abnormal mean to be uniform on $(-0.7, -0.3) \cup (0.3, 0.7)$. Our model for the LOS distribution were negative binomials, with MCEM used to estimate the hyper-parameters of these distributions.

From Table 3 it can be seen that BARD still outperforms PASS especially in regards to accuracy and the number of false positives. The performance of BARD also shows that it is robust to a misspecification of both the LOS distributions and the distribution from which μ was drawn from as we kept the prior in BARD the same. The performance of both methods was impacted by the decreasing mean of the Normal distributions from which μ was drawn as more of them became close to zero and thus abnormal segments became indistinguishable from normal segments.

5.2 Simulated CNV Data

We now make use of the CNV data presented in the Section 1, to obtain a more realistic model to simulate data from. We used the PASS method to initially segment one

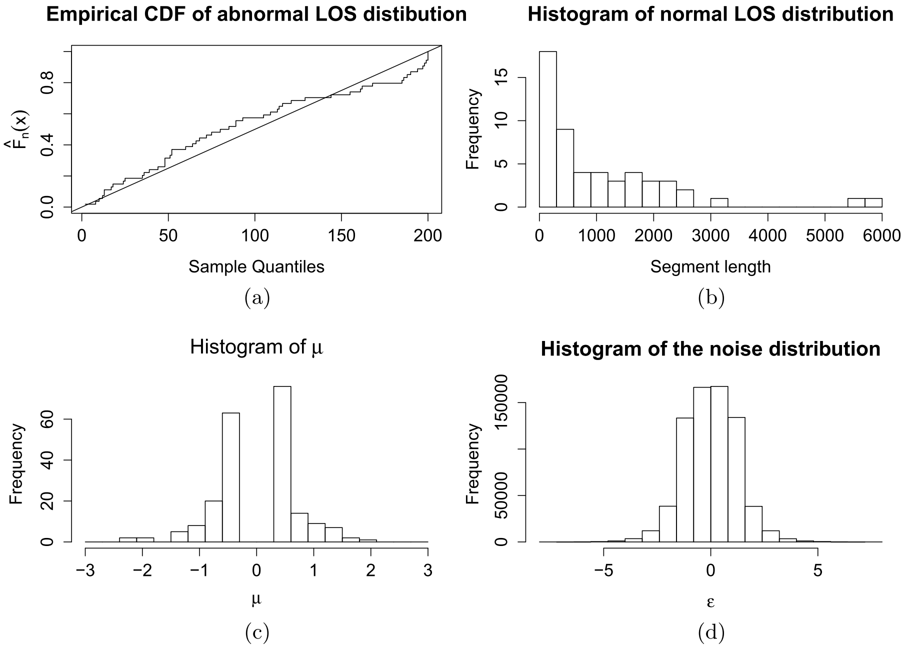
μ	Method	Proportion detected	Accuracy	Number of False positives
$N(0.8, 0.4^2)$	PASS	0.81 (0.78,0.82)	0.065 (0.056,0.068)	1.26 (1.15,1.41)
	BARD	0.85 (0.82,0.86)	0.055 (0.048,0.059)	0.04 (0.02,0.07)
$N(0.7, 0.4^2)$	PASS	0.77 (0.74,0.78)	0.076 (0.069,0.084)	1.11 (1.05,1.33)
	BARD	0.80 (0.78,0.82)	0.066 (0.060,0.073)	0.02 (0.01,0.07)
$N(0.6, 0.4^2)$	PASS	0.69 (0.66,0.71)	0.086 (0.079,0.095)	1.22 (1.08,1.37)
	BARD	0.73 (0.70,0.75)	0.066 (0.061,0.072)	0.06 (0.03,0.09)
$N(0.5, 0.4^2)$	PASS	0.62 (0.60,0.65)	0.10 (0.089,0.11)	1.15 (1.06,1.37)
	BARD	0.65 (0.62,0.68)	0.087 (0.075,0.093)	0.06 (0.02,0.08)
$N(0.4, 0.4^2)$	PASS	0.53 (0.51,0.56)	0.12 (0.10,0.13)	1.07 (0.92,1.22)
	BARD	0.58 (0.55,0.61)	0.093 (0.084,0.10)	0.07 (0.03,0.10)

Table 3: Results based on 200 simulated data sets as we vary the distribution from which μ was simulated from but keeping the prior $\pi(\mu)$ in BARD uniform. The results for each scenario are averages across 200 simulated data sets together with 95% confidence interval in brackets.

replicate of the data, and then analysed this segmentation to obtain information about the LOS distributions and the distributions that generate the data in both normal and abnormal segments.

In Figure 2 we plot some of the empirical data from the segmentation given by PASS. To simulate data sets we either fitted distributions to these quantities or sampled from their empirical distributions. Firstly if we consider the two LOS distributions then for normal segments, see Figure 2b, we found that a geometric distribution fitted the data well. For the abnormal LOS distribution we took a discrete uniform distribution on $\{1, 2, \ldots, 200\}$. This was partly due to us having specified a maximum abnormal segment length of 200 in the PASS method but is potentially realistic in practice as abnormal segments longer than 200 time points are unlikely to occur. To support this choice we plot the empirical cdf of the ordered data and a straight line which are the quantiles of the uniform distribution we propose. We can see that although the fit is not perfect, this is probably due to the small sample size.

Now consider the distributions that generate the actual observations, we can think of these in two parts, one of them being a distribution for the "noise" in normal segments (Figure 2d) and then the mean shift parameter for the abnormal segments (Figure 2c). Up until now we have taken this noise distribution to be standard Normal, however the data suggests that in reality it has heavier tails than the Normal distribution. We found that a t-distribution with 15 degrees of freedom was a better fit to the data so we simulated from this for the noise distribution. For the mean shift parameter μ we took abnormal segments found by the PASS method and looked at the means of each of the dimensions and took the affected dimensions only, this gave the histogram in Figure 2c. In the study we simulated μ from this empirical distribution.



(a) (b)

(c) (d)

Figure 2: Empirical distribution of features of the optimal segmentation of CNV data obtained using the PASS method. (a) QQ-plot of length (measured in number of observations) of abnormal segments against a Uniform distribution on $\{1, 2, \ldots, 200\}$; (b) histogram of length (measured in number of observations) of normal segments; (c) histogram of estimated mean for abnormal segments; and (d) histogram of residuals.

Each simulated data set has length of approximately $n = 20,000$ and dimension $d = 50$. We also varied the proportion of affected dimensions between 4% and 6%. The robustness of BARD to the choice of prior for μ introduced in Section 4, where $|\mu|$ is uniform on an interval (a, b), was also investigated. We simulated 40 of these data sets for each of the scenarios and used both methods to segment them, results are given in Table 4.

We can see that the proportion of correct segments identified is decreased in both methods, this is most likely due to the non-Normally distributed noise present. However the two methods report a very different number of false positives. The performance of BARD is encouraging as it gives many fewer false positives than PASS even with heavier tailed observations than the standard Gaussian case for all choices of a. The results for BARD are similar for different values of $a \leq 0.3$, but do deteriorate slightly for $a = 0.6$. This is likely to be due to a loss of power in detecting abnormal segments with whose change in mean is less than 0.6.

We also vary the second parameter, b, in the prior for μ. We fix $a = 0.3$ and the number of dimensions as $d = 50$. These figures are reported in Table 5 and show that our procedure is relatively robust to the choice of b. For an extreme choice of prior with

% of dim. affected	Method	Proportion detected	Accuracy	Number of False positives
4%	PASS	0.55 (0.51,0.59)	0.071 (0.061,0.081)	1.23 (0.95,1.53)
	BARD $a = 0$	0.65 (0.62,0.70)	0.064 (0.056,0.074)	0.23 (0.10,0.38)
	BARD $a = 0.15$	0.65 (0.61,0.69)	0.065 (0.056,0.074)	0.23 (0.10,0.38)
	BARD $a = 0.3$	0.65 (0.61,0.69)	0.064 (0.055,0.072)	0.2 (0.08,0.35)
	BARD $a = 0.6$	0.55 (0.50,0.59)	0.070 (0.060,0.081)	0.13 (0.03,0.23)
6%	PASS	0.64 (0.61,0.68)	0.068 (0.062,0.074)	1.38 (0.98,1.80)
	BARD $a = 0$	0.72 (0.69,0.75)	0.058 (0.053,0.064)	0.23 (0.10,0.35)
	BARD $a = 0.15$	0.71 (0.68,0.74)	0.059 (0.053,0.065)	0.2 (0.08,0.35)
	BARD $a = 0.3$	0.70 (0.67,0.73)	0.054 (0.049,0.060)	0.1 (0.00,0.20)
	BARD $a = 0.6$	0.63 (0.59,0.66)	0.071 (0.061,0.081)	0.05 (0.00,0.13)

Table 4: Results based on 40 simulated data sets for two scenarios where the proportion of dimensions affected for each abnormal segment varied between 4% and 6% (of the total number of dimensions $d = 50$). The prior for $|\mu|$ assumed by BARD is uniform on $(a, 0.7)$ The results for each case are averages across simulated data sets together with 95% confidence interval in brackets.

$b = 10$ then we can see that the PASS method outperforms BARD on the proportion of segments detected. This highlights that although our method is robust to choices in priors, to ensure the best performance of our method sensible choices need to be made.

b	Method	Proportion detected	Accuracy	Number of False positives
-	PASS	0.55 (0.51,0.59)	0.071 (0.061,0.081)	1.23 (0.95,1.53)
0.5	BARD	0.64 (0.60,0.68)	0.063 (0.055,0.072)	0.13 (0.03,0.23)
1	BARD	0.64 (0.61,0.69)	0.063 (0.055,0.073)	0.3 (0.15,0.48)
2	BARD	0.63 (0.59,0.66)	0.069 (0.059,0.081)	0.2 (0.08,0.35)
4	BARD	0.61 (0.57,0.64)	0.067 (0.057,0.079)	0.13 (0.03,0.25)
10	BARD	0.52 (0.48,0.55)	0.071 (0.060,0.082)	0.10 (0.03,0.20)

Table 5: Results based on 40 simulated data sets for each scenario where the proportion of dimensions affected for each abnormal segment was fixed at 4% and the number of dimensions $d = 50$. The prior for $|\mu|$ used by BARD was $(0.3, b)$. The results for each case are averages across simulated data sets together with 95% confidence interval in brackets.

We also looked at the effect of varying the dimension d of the data, but keeping the proportion of affected dimensions the same. The results can be seen in Table 6, these indicate that both the proportion of abnormal segments detected and the accuracy improve as d is increased, due to the extra information with larger d. However the number of false positives gets worse for both methods, as with larger d there is more chance for some dimensions to show evidence for abnormality within normal regions.

The final parameter we investigate is γ which is instrumental in getting an explicit segmentation from the BARD method using the loss function in (10). If γ is small

d	Method	Proportion detected	Accuracy	Number of False positives
50	PASS	0.55 (0.51,0.59)	0.071 (0.061,0.081)	1.23 (0.95,1.53)
	BARD	0.65 (0.61,0.69)	0.064 (0.055,0.072)	0.2 (0.08,0.35)
100	PASS	0.65 (0.62,0.68)	0.063 (0.056,0.070)	2.1 (1.68,2.58)
	BARD	0.73 (0.70,0.76)	0.051 (0.045,0.059)	0.35 (0.18,0.58)
200	PASS	0.75 (0.72,0.78)	0.055 (0.049,0.062)	3.1 (2.60,3.60)
	BARD	0.85 (0.84,0.87)	0.038 (0.034,0.043)	1.4 (1.05,1.80)

Table 6: Results based on 40 simulated data sets for each scenario where the proportion of dimensions affected for each abnormal segment was fixed at 4% and the number of dimensions d was varied from 50 to 200. The results for each case are averages across simulated data sets together with 95% confidence interval in brackets.

then more evidence is needed for a time point to be classified as abnormal so generally the smaller γ is the smaller the proportion of true positives will be, however the mean number of false positives detected would be large. The reverse is true as γ is increased. We can see from Table 7 that all values of γ perform similarly.

γ	Method	Proportion detected	Accuracy	Number of False positives
-	PASS	0.55 (0.51,0.59)	0.071 (0.061,0.081)	1.23 (0.95,1.53)
1/4	BARD	0.64 (0.60,0.68)	0.064 (0.057,0.073)	0.2 (0.05,0.38)
1/3	BARD	0.65 (0.61,0.69)	0.064 (0.055,0.072)	0.2 (0.08,0.35)
1/2	BARD	0.66 (0.62,0.69)	0.063 (0.054,0.072)	0.28 (0.13,0.45)
2/3	BARD	0.67 (0.63,0.71)	0.063 (0.055,0.073)	0.35 (0.18,0.53)
3/4	BARD	0.67 (0.63,0.71)	0.061 (0.053,0.072)	0.4 (0.23,0.60)
1	BARD	0.68 (0.64,0.71)	0.065 (0.056,0.075)	0.48 (0.28,0.68)

Table 7: Results based on 40 simulated data sets for each scenario where the proportion of dimensions affected for each abnormal segment was fixed at 4%, the number of dimensions $d = 50$ and values of $a = 0.3$ and $b = 0.7$ in the split prior. The parameter γ was varied in the loss function (10). The results for each case are averages across simulated data sets together with 95% confidence interval in brackets.

BARD also allows us to get an estimate of the uncertainty in the position of abnormal segments as from the posterior we can get the probability of each time point belonging to an abnormal segment. If we bin these probabilities into intervals and then find the proportion of these points that are actually abnormal we can obtain a calibration plot Figure 3. We can see from this that the model seems to be well calibrated.

5.3 Analysis of CNV Data

We now apply our method to CNV data from Pinto et al. (2011), a subset of which was presented in Section 1 and was used to construct a model for the simulated data in Section 5.2.

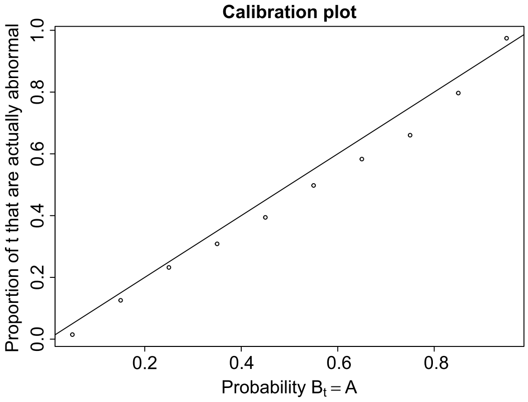


Figure 3: All the time points t for which the posterior probability lies in a certain interval plotted against the proportion of times t lies in an abnormal segment.

Pinto et al. (2011) undertook a detailed study of the different technologies (platforms) used to obtain the measurements and many of the algorithms currently used to call CNV's. We chose to analyse data from the Nimblegen 2.1M platform and from chromosomes 6 and 16. For both chromosomes we have three replicate data sets, each consisting of measurements from six genomes. We preprocessed the data to remove experimental artifacts, using the method described in Siegmund et al. (2011), before analysing it. The data from chromosome 16 consisted of 59,590 measurements, and the data from chromosome 6 consisted of 126,695 measurements, for each genome.

Firstly we ran the PASS method on just the first replicate of the data from chromosome 16 and found the most significant segments. Doing this enables us to get an estimate of the parameters for the LOS distributions to use in the Bayesian method without having to do any parameter inference. The maximum length of segment we searched over was 200 (measured in observations not base pairs) as this is greater than the largest CNV we would expect to find. This gave parameters that suggested a geometric distribution for the length of normal segments $S_N \sim \text{Geom}(0.0007)$ and the following Negative Binomial distribution for abnormal segments $S_N \sim \text{NBinom}(2, 0.1)$. We used the same split uniform prior for μ as we did in Section 5.2 namely one with equal density on the set $(-0.7, -0.3) \cup (0.3, 0.7)$ and zero elsewhere. We justified the use of this form of prior which excludes values close to zero in Section 2.2 and it was shown to perform well on some realistically simulated data in Section 5.2.

For both chromosomes we analysed the three replicates separately. Ideally we should infer exactly the same segmentation for each of the replicate data sets. Due to the large amount of noise present in the data this does not happen. However we would expect that a "better" method would be more consistent across the three replicates, and we use the consistency of the inferred segmentations across the replicates as a measure of accuracy.

We can also use data from the HapMap project to validate some of the CNV's

Truth		PASS			BARD		
Start	Length	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
2619669	62144	-	-	-	-	✓	✓
21422575	76266	✓	✓	✓	✓	✓	✓
32165010	456897	✓	✓	✓	✓	✓	✓
34328205	286367	✓	✓	✓	✓	✓	✓
54351338	28607	✓	✓	-	✓	✓	✓
70644511	21083	-	✓	✓	-	✓	✓

Table 8: Known CNV's from HapMap found by either method when analysing different replicates of data from chromosome 16. Ticks indicate whether the particular segment was detected or not.

Truth		PASS			Bayesian		
Start	Length	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
202353	37484	-	-	-	✓	✓	-
243700	80315	✓	✓	✓	✓	✓	✓
29945167	12079	✓	✓	-	-	-	-
31388080	61239	✓	-	-	✓	-	-
32562253	117686	✓	-	✓	-	-	-
32605094	74845	-	✓	-	✓	✓	✓
32717276	22702	✓	-	-	✓	✓	✓
74648953	9185	✓	✓	-	✓	✓	✓
77073620	10881	-	✓	-	✓	✓	✓
77155307	781	-	-	-	✓	-	-
77496587	12936	-	-	-	✓	✓	✓
78936990	18244	✓	✓	✓	✓	✓	✓
103844669	24085	✓	✓	✓	✓	✓	✓
126225385	3084	✓	✓	-	-	-	✓
139645437	3392	-	-	-	✓	-	-
165647807	4111	-	-	-	✓	-	✓

Table 9: Known CNV's from HapMap found by either method when analysing different replicates of data from chromosome 6. Ticks indicate whether the particular segment was detected or not.

we found to those known experimentally or which have been called by other authors. A list containing these known CNV's by chromosome and sample can be found at http://hapmap.ncbi.nlm.nih.gov/. These validated segments suggest that about 1% of chromosome 16 is abnormal.

To make comparisons between BARD and PASS fair we implemented both of these methods so that they identified the same proportion, 4%, of the chromosome as being abnormal. For BARD this involved choosing γ in the loss function (10) appropriately and for PASS selecting the most significant segments that give us a total of 4% abnormal time points. We then tested these against the validated CNV's.

Chromosome	Method	Rep 1 v 2	Rep 1 v 3	Rep 2 v 3
6	PASS	0.474	0.709	0.522
	BARD	0.495	0.457	0.416
16	PASS	0.478	0.507	0.388
	BARD	0.426	0.467	0.682

Table 10: The average consistency measured using the dissimilarity measure for found CNV's between replicates and methods. A lower value indicates the inferred segmentations for the two replicates were more similar.

The results for chromosome 16 are contained in Tables 8 and 10; and those for chromosome 6 in Tables 9 and 10. Tables 8 and 9 list the known CNV regions that were detected by one or both methods for at least one replicate, whilst Table 10 gives summaries of the consistency of the inferred segmentations across replicates.

The results show that BARD is more successful at detecting known CNV regions than PASS. In total BARD found 6 CNV regions on chromosome 16 for at least one replicate, and 14 for chromosome 6, while PASS managed 5 and 11 respectively. For the measures of consistency across the different replicates, shown in Table 10, BARD performed better for 4 of the 6 pairs.

6 Discussion

In this paper we have developed novel methodology to detect abnormal regions in multiple time series. Firstly we developed a general model for this type of problem including length of stay distributions and marginal likelihoods for normal and abnormal segments. We then derived recursions that could be used to calculate the posterior of interest and showed how to obtain iid samples from an accurate approximation to this posterior in a way that scales linearly with the length of series.

The resulting algorithm, BARD, was then compared in several simulation studies and some real data to another competing method PASS. These results showed that BARD was consistently more accurate than the PASS benchmark on several important criteria for all of the data sets we considered. Furthermore, being able to accurately and efficiently perform Bayesian inference for large and high dimensional data sets of this type allows us to quantify uncertainty in the location of abnormal segments. Before this with other methods such as PASS this quantification of uncertainty has not been possible.

Whilst we have focused on a specific model of changes in mean from some baseline level, our method could easily be adapted to any model which specifies some normal behaviour and abnormal behaviour. The only restrictions we place on this is the ability to calculate marginal likelihoods for both types of segment. The main computational bottleneck would be in the calculation of the abnormal marginal likelihoods as this involves integration over a prior for the parameter(s) which cannot be done analytically, and for higher dimensional parameters would be computationally intensive. For example, our approach can trivially be extended to allow different but known variances for each time-series. To allow each abnormal segment to have its own variance as well as mean is possible, but would involve extra computation, as a 2-dimensional integral would be needed to calculate the marginal likelihoods for abnormal segments.

R code to run the BARD method is available at the first authors website. `http://www.lancaster.ac.uk/pg/bardwell/Work.html`. The real CNV data we analysed in Section 5.3 is available publicly and can be downloaded from the GEO accession website `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25893`.

Supplementary Material

Supplementary Material of "Bayesian detection of abnormal segments in multiple time series" (DOI: 10.1214/16-BA998SUPP; .pdf).

References

Bardwell, L. and Fearnhead, P. (2016). "Supplementary Material of "Bayesian detection of abnormal segments in multiple time series"." *Bayesian Analysis*. doi: `http://dx.doi.org/10.1214/16-BA998SUPP`. 203

Barry, D. and Hartigan, J. A. (1992). "Product partition models for change point problems." *The Annals of Statistics*, 20(1): 260–279. `http://www.jstor.org/stable/2242159`. MR1150343. doi: `http://dx.doi.org/10.1214/aos/1176348521`. 194

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. Springer series in statistics. New York, NY [u.a.]: Springer, 2. ed edition. `http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+027440176&sourceid=fbw_bibsonomy`. MR0804611. doi: `http://dx.doi.org/10.1007/978-1-4757-4286-2`. 202

Carter, C. K. and Kohn, R. (1994). "On Gibbs sampling for state space models." *Biometrika*, 81(3): 541–553. MR1311096. doi: `http://dx.doi.org/10.1093/biomet/81.3.541`. 202

Cox, D. (1962). *Renewal Theory*. Methuen's monographs on applied probability and statistics. Methuen. `http://books.google.co.uk/books?id=0VxRAAAAMAAJ`. MR0153061. 196

Fearnhead, P. (2006). "Exact and efficient Bayesian inference for multiple changepoint problems." *Statistics and Computing*, 16(2): 203–213. `http://www.springerlink.com/content/51j72n74761l011q/`. MR2227396. doi: `http://dx.doi.org/10.1007/s11222-006-8450-8`. 194

Fearnhead, P. and Liu, Z. (2007). "On-line inference for multiple changepoint problems." *Journal of the Royal Statistical Society B*, 69: 589–605. `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2007.00601.x/abstract`. MR2370070. doi: `http://dx.doi.org/10.1111/j.1467-9868.2007.00601.x`. 201

Fearnhead, P. and Vasileiou, D. (2009). "Bayesian analysis of isochores." *Journal of the American Statistical Association*, 104(485): 132–141. http://pubs.amstat.org/doi/abs/10.1198/jasa.2009.0009?journalCode=jasa. MR2663038. doi: http://dx.doi.org/10.1198/jasa.2009.0009. 194, 200

Frick, K., Munk, A., and Sieling, H. (2014). "Multiscale change point inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3): 495–580. http://dx.doi.org/10.1111/rssb.12047. MR3210728. doi: http://dx.doi.org/10.1111/rssb.12047. 194

Galeano, P., Peña, D., and Tsay, R. S. (2006). "Outlier detection in multivariate time series by projection pursuit." *Journal of the American Statistical Association*, 101(474): 654–669. http://www.jstor.org/stable/27590725. MR2256180. doi: http://dx.doi.org/10.1198/016214505000001131. 194

Jandhyala, V., Fotopoulos, S., MacNeill, I., and Liu, P. (2013). "Inference for single and multiple change-points in time series." *Journal of Time Series Analysis*. MR3070866. doi: http://dx.doi.org/10.1111/jtsa.12035. 194

Jeng, X. J., Cai, T. T., and Li, H. (2013). "Simultaneous discovery of rare and common segment variants." *Biometrika*, 100(1): 157–172. http://www.biomedsearch.com/nih/Simultaneous-Discovery-Rare-Common-Segment/23825436.html. MR3034330. doi: http://dx.doi.org/10.1093/biomet/ass059. 194, 199, 200, 202, 203, 206

Jin, J. (2004). *Detecting a Target in Very Noisy Data from Multiple Looks*, volume 45 of *Lecture Notes – Monograph Series*, 255–286. Beachwood, Ohio, USA: Institute of Mathematical Statistics. http://dx.doi.org/10.1214/lnms/1196285396. MR2126903. doi: http://dx.doi.org/10.1214/lnms/1196285396. 193

Kulkarni, V. (2012). *Introduction to Modeling and Analysis of Stochastic Systems*. Springer Texts in Statistics. Springer London, Limited. http://books.google.co.uk/books?id=2EeGkQEACAAJ. MR2743408. doi: http://dx.doi.org/10.1007/978-1-4419-1772-0. 197

Levine, R. A. and Casella, G. (2001). "Implementations of the Monte Carlo EM Algorithm." *Journal of Computational and Graphical Statistics*, 10(3): 422–439. MR1939033. doi: http://dx.doi.org/10.1198/106186001317115045. 202

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). "Circular binary segmentation for the analysis of array based DNA copy number data." *Biostatistics*, 5(4): 557–572. http://biostatistics.oxfordjournals.org/content/5/4/557.abstract. 194

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A. C., Thiruvahindrapuram, B., MacDonald, J. R., Mills, R., Prasad, A., Noonan, K., Gribble, S., Prigmore, E., Donahoe, P. K., Smith, R. S., Park, J. H., Hurles, M. E., Carter, N. P., Lee, C., Scherer, S. W., and Feuk, L. (2011). "Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants." *Nature Biotechnology*, 29(6): 512–521. 212

Qu, G., Hariri, S., and Yousif, M. (2005). “Multivariate statistical analysis for network attacks detection.” In *The 3rd ACS/IEEE International Conference on Computer Systems and Applications, 2005*, 9. 193

Siegmund, D., Yakir, B., and Zhang, N. R. (2011). “Detecting simultaneous variant intervals in aligned sequences.” *The Annals of Applied Statistics*, 5(2A): 645–668. MR2840169. doi: http://dx.doi.org/10.1214/10-AOAS400. 194, 213

Spiegel, S., Gaebler, J., Lommatzsch, A., De Luca, E., and Albayrak, S. (2011). “Pattern recognition and classification for multivariate time series.” In *Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data*, SensorKDD'11, 34–42. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2003653.2003657. 193

Tsay, R. S., Peña, D., and Pankratz, A. E. (2000). “Outliers in multivariate time series.” *Biometrika*, 87(4): 789–804. http://biomet.oxfordjournals.org/content/87/4/789.abstract. MR1813975. doi: http://dx.doi.org/10.1093/biomet/87.4.789. 194

Wyse, J., Friel, N., and Rue, H. (2011). “Approximate simulation-free Bayesian inference for multiple changepoint models with dependence within segments.” *Bayesian Analysis*, 6(4): 501–528. MR2869956. 194

Yau, C. and Holmes, C. C. (2010). “A decision theoretic approach for segmental classification using Hidden Markov models.” http://arxiv.org/abs/1007.4532 202

Zhang, N. (2010). “DNA copy number profiling in normal and tumor genomes.” In Feng, J., Fu, W., and Sun, F. (eds.), *Frontiers in Computational and Systems Biology*, volume 15 of *Computational Biology*, 259–281. Springer London. http://dx.doi.org/10.1007/978-1-84996-196-7_14. 194

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z. (2010). “Detecting simultaneous changepoints in multiple sequences.” *Biometrika*, 97(3): 631–645. http://ideas.repec.org/a/oup/biomet/v97y2010i3p631-645.html. MR2672488. doi: http://dx.doi.org/10.1093/biomet/asq025. 194

Acknowledgments

We thank Idris Eckley for helpful comments and discussions. This research was supported by EPSRC grant EP/K014463/1. Bardwell gratefully acknowledges the support of British Telecom and the EPSRC funded EP/H023151/1 STOR-i Centre for Doctoral Training.