Bayesian detection of abnormal segments in multiple time series

Lawrence Bardwell and Paul Fearnhead

December 18, 2014

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, 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 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 \( \log_2 \) 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 the first two individuals. This can see by the raised means (indicated by the red lines) in these two 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 point estimate of the which are the abnormal regions. Here we introduce a Bayesian approach to detecting abnormal regions. This is able to both give point estimates of the number and location of the abnormal regions, and also to 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 CNV application.
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 indi-
viduals in red. For this CNV only the first two individuals (NA10851 and NA12239) are
affected.

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 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, \( \mathbf{Y}_{1:n} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n) \) where \( \mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,d})^T \). 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, \( p(y_{1:n}|x_{1:n}) \). 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

\[
p(x_{1:n}|y_{1:n}) \propto p(x_{1:n}, y_{1:n}) = p(x_{1:n})p(y_{1:n}|x_{1:n}). \tag{2.1}
\]

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 \( \pi_N \) and \( \pi_A \) 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 \( C_t \in \{0, \ldots, t-1\} \) 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
\[
p(x_{1:n}) = p(c_{1:n}, b_{1:n}) = \Pr(C_1 = c_1, B_1 = b_1) \prod_{i=1}^{n-1} \Pr(C_{i+1} = c_{i+1}, B_{i+1} = b_{i+1}|C_i = c_i, B_i = b_i).
\] (2.2)

The decomposition in (2.2) 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 \( i = 1, \ldots, t - 1 \) we have

\[
\Pr(C_{t+1} = j, B_{t+1} = k|C_t = i, B_t = N) = \begin{cases} 
\frac{1-G_N(t-i)}{1-G_N(t-i-1)} & \text{if } j = i \text{ and } k = N, \\
\frac{G_N(t-i)-G_N(t-i-1)}{1-G_N(t-i)} & \text{if } j = t \text{ and } k = A, \\
0 & \text{otherwise},
\end{cases}
\]

\[
\Pr(C_{t+1} = j, B_{t+1} = k|C_t = i, B_t = A) = \begin{cases} 
\frac{1-G_A(t-i)}{1-G_A(t-i-1)} & \text{if } j = i \text{ and } k = A, \\
\frac{G_A(t-i)-G_A(t-i-1)}{1-G_A(t-i)} & \text{if } j = t \text{ and } k = A, \\
\pi_A & \text{if } j = t \text{ and } k = N, \\
\pi_N & \text{otherwise}.
\end{cases}
\] (2.3)

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, \( p(y_{1:n}|x_{1:n}) \).

To make this model tractable we 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. Thus when we condition on $C_t$ and $B_t$ the likelihood for the first $t$ observations factorises as follows

$$p(y_{1:t}|C_t = j, B_t) = p(y_{1:j}|C_t = j, B_t)p(y_{j+1:t}|C_t = j, B_t).$$  \hspace{1cm} (2.4)$$

The second term in equation (2.4) is the marginal likelihood of the data, $Y_{j+1:t}$, given it comes from a segment that has type $B_t$. We introduce the following notation for these segment marginal likelihoods, where for $s \geq t$,

$$P_N(t, s) = \Pr(y_{t:s}|C_s = t - 1, B_s = N),$$

$$P_A(t, s) = \Pr(y_{t:s}|C_s = t - 1, B_s = A),$$

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 (2.6) due to the conditional independence assumption on the segments

$$p(y_{1:n}|x_{1:n}) = \prod_{t=1}^{n} p(y_t|x_{1:n}, y_{1:(t-1)}) = \prod_{t=1}^{n} p(y_t|C_t, B_t, y_{(C_t+1):(t-1)}).$$ \hspace{1cm} (2.6)$$

The terms on the right-hand side of equation (2.6) can then be written in terms of the segment marginal likelihoods

$$p(y_t|C_t, B_t, y_{(C_t+1):(t-1)}) = \frac{P_B(C_t + 1, t)}{P_B(C_t + 1, t - 1)}.$$ \hspace{1cm} (2.7)$$

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

### 2.2.1 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, $D$, and these realisations are independent over both time and dimension. Denote the density function of the distribution $D$ as $f_D(\cdot)$. We can write down the segment marginal likelihood as

$$P_N(t, s) = \prod_{k=1}^{d} \prod_{i=t}^{s} f_D(y_{i,k}).$$ \hspace{1cm} (2.8)$$

### 2.2.2 Model for data in abnormal segments

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

6
Our model for which dimensions have data drawn from $P_\theta$ 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 $\theta$, the likelihood of the data associated with the $k$th dimension is

$$p_k \prod_{i=t}^{s} f_P(y_{i,k}|\theta) + (1 - p_k) \prod_{i=t}^{s} f_D(y_{i,k}).$$

Thus by independence over dimension

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

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

$$P_A(t,s) = \int p(\mathbf{y}_{t:s}|\theta)\pi(\theta)\,d\theta. \quad (2.9)$$

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

### 2.2.3 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 $\mu$ depending on whether we are in a normal or abnormal segment. This model also underpins the simulation studies that we present in Section 4.

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 $\sigma^2$ is constant and known (in practice we would be able to estimate it from the data).

Having specified these two distributions we then need to calculate marginal likelihoods for normal and abnormal segments given by equations (2.8) and (2.9) respectively. Calculating the marginal likelihood for a normal segment is simple because of independence over time and dimension as shown in equation (2.8). 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 $\mu$ to be uniform on a region that excludes values of $\mu$ 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 Vasiliou (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 up to time $t$, $p(x_t | y_{1:t}) = p(c_t, b_t | y_{1:t})$, 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, $p(x_{1:n} | y_{1:n})$.

To find these filtering distribution we develop a set of recursions that enable us to calculate $p(c_{t+1}, b_{t+1} | y_{1:t+1})$ in terms of $p(c_t, b_t | y_{1:t})$. 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 $k \in \{N, A\}$,

$$
p(C_{t+1} = j, B_{t+1} = k | y_{1:(t+1)}) \propto p(y_{t+1} | y_{1:t}, C_{t+1} = j, B_{t+1} = k) p(C_{t+1} = j, B_{t+1} = k | y_{1:t})
$$

$$
= \left( \frac{P_k(j + 1, t + 1)}{P_k(j + 1, t)} \right) \Pr(C_{t+1} = j, B_{t+1} = k | C_t = j, B_t = k) p(C_t = j, B_t = k | y_{1:t}),
$$

where the first term in the last expression is the conditional likelihood from equation (2.7). 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 $C_{t+1} = t$,

$$
p(C_{t+1} = t, B_{t+1} = k | y_{1:t}) = \sum_{i=0}^{t-1} \sum_{l \in \{N, A\}} p(C_t = i, B_t = l | y_{1:t}) \Pr(C_{t+1} = t, B_{t+1} = k | C_t = i, B_t = l).
$$

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

$$
p(C_{t+1} = t, B_{t+1} = k | y_{1:(t+1)}) \propto P_k(t + 1, t + 1) \sum_{i=0}^{t-1} \sum_{l \in \{N, A\}} p(C_t = i, B_t = l | y_{1:t}) \Pr(C_{t+1} = t, B_{t+1} = k | C_t = i, B_t = l).
$$

These recursions are initialised by $p(C_1 = 0, B_1 = k | y_1) \propto \Pr(B_1 = k) P_k(1, 1)$ for $k \in \{N, A\}$.

3.2 Approximate Inference

The support of the filtering distribution $p(c_t, b_t | y_{1:t})$ has $2t$ points. Hence, calculating $p(c_t, b_t | y_{1:t})$ 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 \( \text{[Fearnhead and Liu 2007]} \) 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 \( \alpha \). 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 \( \alpha \).

### 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. First we simulate \( X_n = (C_n, B_n) \) from the final filtering distribution \( p(c_n, b_n|y_{1:n}) \). 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 \( Y_{1:n} \),

\[
p(c_t, b_t|C_{t+1} = t, B_{t+1}, y_{1:n}) \\
\propto p(c_t, b_t, C_{t+1} = t, B_{t+1}, y_{1:n}) \\
= p(c_t, b_t) \Pr(C_{t+1} = t, B_{t+1}|C_t, B_t)p(y_{1:n}|C_t, B_t, C_{t+1} = t, B_{t+1}) \\
\propto p(c_t, b_t) \Pr(C_{t+1} = t, B_{t+1}|C_t, B_t)p(y_{1:t}|C_t, B_t) \\
\propto p(c_t, b_t|y_{1:t}) \Pr(C_{t+1} = t, B_{t+1}|C_t, B_t).
\]

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 \( (2.1) \) depends upon a vector of hyper-parameters which we now label as \( \Psi \). Specifically \( \Psi \) contains the parameters for the LOS distributions for the two differing types of segments which determine the cdf’s \( G_N(\cdot) \) and \( G_A(\cdot) \).

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 \( \text{[Levine and Casella 2001]} \). Although convergence of the hyper-parameters is quite rapid in the examples we look at in Section \([4]\) 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 4.

3.5 Estimating a Segmentation

We have described how to calculate the posterior density \( p(x_{1:n}|y_{1:n}) \) 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:

\[
L(\tilde{b}_t|b_t) = \begin{cases} 
1 & \text{if } \tilde{b}_t = A \text{ and } b_t = N \\
\gamma & \text{if } \tilde{b}_t = N \text{ and } b_t = A \\
0 & \text{otherwise}
\end{cases} \quad (3.1)
\]

The inclusion of \( \gamma \) 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 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 2.2.3, 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 \( k = \{1, 2, \ldots, d\} \). 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 \( \mu \) 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 \( \alpha \) gave a good trade off between accuracy and computational cost.

To get an explicit segmentation from BARD we use the asymmetric loss function (3.1) 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)](http://example.com). 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 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. 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 $D_k$ defined in [Jeng et al. (2013)](http://example.com).

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

\[
D_k = \min_{I_j \in \hat{I}} \left\{ 1 - \frac{|\hat{I}_j \cap I_k|}{\sqrt{|\hat{I}_j||I_k|}} \right\}
\]  

$D_k \in [0, 1]$, 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 $k^{th}$ segment, i.e. it hasn’t been detected. Smaller values of $D$ indicate a greater overlap.

### 4.1 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

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

Two different distributions were used to generate the altered means for the affected dimensions and we also varied $\pi_N$ (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 $\pi_N$. 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.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\pi_N$</th>
<th>Method</th>
<th>Proportion detected</th>
<th>Accuracy</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U(0.3,0.7)$</td>
<td>0.5</td>
<td>PASS</td>
<td>(0.66,0.70)</td>
<td>(0.11,0.13)</td>
<td>(0.68,0.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD</td>
<td>(0.87,0.89)</td>
<td>(0.070,0.084)</td>
<td>(0.04,0.12)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>PASS</td>
<td>(0.66,0.71)</td>
<td>(0.12,0.14)</td>
<td>(0.90,1.19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD</td>
<td>(0.75,0.78)</td>
<td>(0.081,0.093)</td>
<td>(0.03,0.09)</td>
</tr>
<tr>
<td>$U(0.5,0.9)$</td>
<td>0.5</td>
<td>PASS</td>
<td>(0.91,0.93)</td>
<td>(0.070,0.077)</td>
<td>(0.93,1.22)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD</td>
<td>(0.98,0.99)</td>
<td>(0.035,0.042)</td>
<td>(0.01,0.06)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>PASS</td>
<td>(0.93,0.95)</td>
<td>(0.069,0.076)</td>
<td>(0.88,1.17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD</td>
<td>(0.95,0.97)</td>
<td>(0.040,0.045)</td>
<td>(0.00,0.04)</td>
</tr>
</tbody>
</table>

Table 1: Scenarios differed in the prior for $\mu$ and the value of $\pi_N$ used to simulate the data. In BARD these same priors were used for the analysis of the data. The results are based on 200 simulated data sets for each scenario and the intervals given are 95% confidence intervals calculated using 1000 bootstrap replicates.

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.

<table>
<thead>
<tr>
<th>$p_k$</th>
<th>Proportion detected</th>
<th>Accuracy</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{200}$</td>
<td>(0.63,0.67)</td>
<td>(0.086,0.10)</td>
<td>(0.005,0.06)</td>
</tr>
<tr>
<td>$\frac{4}{200}$</td>
<td>(0.74,0.78)</td>
<td>(0.086,0.10)</td>
<td>(0.05,0.12)</td>
</tr>
<tr>
<td>$\frac{8}{200}$</td>
<td>(0.75,0.78)</td>
<td>(0.081,0.093)</td>
<td>(0.03,0.09)</td>
</tr>
<tr>
<td>$\frac{12}{200}$</td>
<td>(0.74,0.78)</td>
<td>(0.083,0.096)</td>
<td>(0.03,0.095)</td>
</tr>
<tr>
<td>$\frac{16}{200}$</td>
<td>(0.72,0.76)</td>
<td>(0.084,0.098)</td>
<td>(0.03,0.09)</td>
</tr>
<tr>
<td>$\frac{20}{200}$</td>
<td>(0.70,0.74)</td>
<td>(0.088,0.102)</td>
<td>(0.02,0.08)</td>
</tr>
</tbody>
</table>

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 $p_k$.

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 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.

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>Method</th>
<th>Proportion detected</th>
<th>Accuracy</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(0.8,0.4^2) )</td>
<td>PASS</td>
<td>(0.78,0.82)</td>
<td>(0.056,0.068)</td>
<td>(1.15,1.41)</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>(0.82,0.86)</td>
<td>(0.048,0.059)</td>
<td>(0.02,0.07)</td>
</tr>
<tr>
<td>( N(0.7,0.4^2) )</td>
<td>PASS</td>
<td>(0.74,0.78)</td>
<td>(0.069,0.084)</td>
<td>(1.05,1.33)</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>(0.78,0.82)</td>
<td>(0.060,0.073)</td>
<td>(0.01,0.07)</td>
</tr>
<tr>
<td>( N(0.6,0.4^2) )</td>
<td>PASS</td>
<td>(0.66,0.71)</td>
<td>(0.079,0.095)</td>
<td>(1.08,1.37)</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>(0.70,0.75)</td>
<td>(0.061,0.072)</td>
<td>(0.03,0.09)</td>
</tr>
<tr>
<td>( N(0.5,0.4^2) )</td>
<td>PASS</td>
<td>(0.60,0.65)</td>
<td>(0.089,0.11)</td>
<td>(1.06,1.37)</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>(0.62,0.68)</td>
<td>(0.075,0.093)</td>
<td>(0.02,0.08)</td>
</tr>
<tr>
<td>( N(0.4,0.4^2) )</td>
<td>PASS</td>
<td>(0.51,0.56)</td>
<td>(0.10,0.13)</td>
<td>(0.92,1.22)</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>(0.55,0.61)</td>
<td>(0.084,0.10)</td>
<td>(0.03,0.10)</td>
</tr>
</tbody>
</table>

Table 3: Results based on 200 simulated data sets as we vary the distribution from which \( \mu \) was simulated from but keeping the prior \( \pi(\mu) \) in BARD uniform. 95\% confidence intervals for the means were calculated using 1000 bootstrap replicates.

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 \( \mu \) 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 \( \mu \) was drawn as more of them became close to zero and thus abnormal segments became indistinguishable from normal segments.

### 4.2 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 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 2a, 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 $\mu$ 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 $\mu$ from this empirical distribution.

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%. We simulated 40 of these data sets for each of the two scenarios and used both methods to segment them, results are given in Table 4.

<table>
<thead>
<tr>
<th>% of affected dimensions</th>
<th>Method</th>
<th>Proportion detected</th>
<th>Accuracy</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>4%</td>
<td>PASS</td>
<td>(0.59, 0.66)</td>
<td>(0.080, 0.10)</td>
<td>(7.15, 9.23)</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>(0.61, 0.69)</td>
<td>(0.055, 0.072)</td>
<td>(0.08, 0.38)</td>
</tr>
<tr>
<td>6%</td>
<td>PASS</td>
<td>(0.64, 0.72)</td>
<td>(0.066, 0.085)</td>
<td>(2.25, 3.03)</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>(0.71, 0.78)</td>
<td>(0.046, 0.060)</td>
<td>(0.10, 0.43)</td>
</tr>
</tbody>
</table>

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 $d = 50$). 95% confidence intervals for the means were calculated using 1000 bootstrap replicates.

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.

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.

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.
4.3 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 4.2 and was used to construct a model for the simulated data in Section 4.2. 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 $\mu$ as we did in Section 4.2, 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 2.2.3 and it was shown to perform well on some realistically simulated data in Section 4.2.

<table>
<thead>
<tr>
<th>Truth</th>
<th>PASS</th>
<th>BARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>Length</td>
<td>Rep 1</td>
</tr>
<tr>
<td>2619669</td>
<td>62144</td>
<td>-</td>
</tr>
<tr>
<td>21422575</td>
<td>76266</td>
<td>✓</td>
</tr>
<tr>
<td>32165010</td>
<td>456897</td>
<td>✓</td>
</tr>
<tr>
<td>34328205</td>
<td>286367</td>
<td>✓</td>
</tr>
<tr>
<td>54351338</td>
<td>28607</td>
<td>✓</td>
</tr>
<tr>
<td>70644511</td>
<td>21083</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5: 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.

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 we found to
Table 6: 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.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Method</th>
<th>Rep 1 v 2</th>
<th>Rep 1 v 3</th>
<th>Rep 2 v 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>PASS</td>
<td>0.474</td>
<td>0.709</td>
<td>0.522</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>0.495</td>
<td>0.457</td>
<td>0.416</td>
</tr>
<tr>
<td>16</td>
<td>PASS</td>
<td>0.478</td>
<td>0.507</td>
<td>0.388</td>
</tr>
<tr>
<td></td>
<td>BARD</td>
<td>0.426</td>
<td>0.467</td>
<td>0.682</td>
</tr>
</tbody>
</table>

Table 7: 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.

Table 6: 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.

<table>
<thead>
<tr>
<th>Truth</th>
<th>PASS</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>Length</td>
<td>Rep 1</td>
</tr>
<tr>
<td>202353</td>
<td>37484</td>
<td>✓</td>
</tr>
<tr>
<td>243700</td>
<td>80315</td>
<td>✓</td>
</tr>
<tr>
<td>29945167</td>
<td>12079</td>
<td>✓</td>
</tr>
<tr>
<td>31388080</td>
<td>61239</td>
<td>✓</td>
</tr>
<tr>
<td>32562253</td>
<td>117686</td>
<td>✓</td>
</tr>
<tr>
<td>32605094</td>
<td>74845</td>
<td>✓</td>
</tr>
<tr>
<td>32717276</td>
<td>22702</td>
<td>✓</td>
</tr>
<tr>
<td>74648953</td>
<td>9185</td>
<td>✓</td>
</tr>
<tr>
<td>77073620</td>
<td>10881</td>
<td>✓</td>
</tr>
<tr>
<td>77155307</td>
<td>781</td>
<td>-</td>
</tr>
<tr>
<td>77496587</td>
<td>12936</td>
<td>-</td>
</tr>
<tr>
<td>78936990</td>
<td>18244</td>
<td>✓</td>
</tr>
<tr>
<td>103844669</td>
<td>24085</td>
<td>✓</td>
</tr>
<tr>
<td>126225385</td>
<td>3084</td>
<td>✓</td>
</tr>
<tr>
<td>139645437</td>
<td>3392</td>
<td>-</td>
</tr>
<tr>
<td>165647807</td>
<td>4111</td>
<td>-</td>
</tr>
</tbody>
</table>

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 $\gamma$ in the loss function (3.1) 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.

The results for chromosome 16 are contained in Tables 5 and 7 and those for chromosome 6 in Tables 6 and 7. Tables 5 and 6 list the known CNV regions that were detected by one or both methods for at least one replicate, whilst Table 7 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 7, BARD performed better for 4 of the 6 pairs.

5 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.

The novelty of our method comes from being able to accurately and efficiently perform Bayesian inference for large and high dimensional data sets of this type thus allowing 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 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 only potential 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.

R code to run the BARD method is available at the first authors website. The real CNV data we analysed in Section 4.3 is available publicly and can be downloaded from the GEO accession website.

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

References


