Published in final edited form as:
Comput Stat Data Anal. 2009 July 1; 53(9): 3314–3323. doi:10.1016/j.csda.2009.02.006.

Non-iterative sampling-based Bayesian methods for identifying changepoints in the sequence of cases of haemolytic uraemic syndrome

Guo-Liang Tian[†,*], **Kai Wang Ng**[†], **Kai-Can Li**[‡], and **Ming Tan**[§]

[†] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, P. R. China

[‡] School of Mathematics and Statistics, Hubei Normal University, 82 Cihu Road, Huangshi City, Hubei 435002, P. R. China

[§] Division of Biostatistics, University of Maryland Greenebaum Cancer Center, MSTF Suite 261, 10 South Pine Street, Baltimore, Maryland 21201, U.S.A

Summary

Diarrhoea-associated *Haemolytic uraemic syndrome* (HUS) is a disease that affects the kidneys and other organs. Motivated by the annual number of cases of HUS collected in Birmingham and Newcastle of England, respectively, from 1970 to 1989, we consider Bayesian changepoint analysis with specific attention to Poisson changepoint models. For changepoint models with unknown number of changepoints, we propose a new non-iterative Bayesian sampling approach (called exact IBF sampling), which completely avoids the problem of convergence and slow convergence associated with iterative *Markov chain Monte Carlo* (MCMC) methods. The idea is to first utilize the sampling *inverse Bayes formula* (IBF) to derive the conditional distribution of the latent data given the observed data, and then to draw iid samples from the complete-data posterior distribution. For the purpose of selecting the appropriate model (or determining the number of changepoints), we develop two alternative formulae to exactly calculate marginal likelihood (or Bayes factor) by using the exact IBF output and the point-wise IBF, respectively. The HUS data are re-analyzed using the proposed methods. Simulations are implemented to validate the performance of the proposed methods.

Keywords

Bayes factor; Changepoint problem; Haemolytic uraemic syndrome; IBF sampling; MCMC; Non-iterative Bayesian approach; Poisson distribution

1. Introduction

Diarrhoea-associated *Haemolytic uraemic syndrome* (HUS) is a disease that affects the kidneys and other organs. It poses a substantial threat to infants and young children as one of the leading causes of both acute and chronic kidney failures. HUS is most common in the warmer months

*Corresponding author's email: gltian@hku.hk.

of the year, following a gastrointestinal illness caused primarily by a particular strain of bacterium, Escherichia Coli O157:H7 (Milford *et al*., 1990). These bacteria (E. Coli O157:H7) produce extremely potent toxins which are the main cause of the symptoms related to the gastrointestinal illness. Table 1 displays the annual number of cases of HUS collected in Birmingham and Newcastle of England, respectively, from 1970 to 1989 (Tarr *et al*., 1989; Henderson and Matthews, 1993). The primary concern is the incidence of HUS and when the frequency of cases increases sharply. In the mean-corrected cumulative sum plot (Figure 1), the annual totals appear to increase abruptly at about 1980 for the Birmingham series and 1976, 1984 for the Newcastle series. Therefore, a changepoint analysis of the data with Poisson models seems to be appropriate.

Changepoint problems (CPPs) are often encountered in medicine and other fields, e.g., economics, finance, psychology, signal processing, industrial system control and geology. Typically, a sequence of data is collected over a period of time, we wish to make inference about the location of one or more points of the sequence at which there is a change in the model. The literature on CPPs is extensive. For Poisson process CPPs, a well-known example concerns British coal-mining disasters from 1851–1962 (originally gathered by Maguire *et al*. (1952) and corrected by Jarrett (1979)). Frequentist investigations appear in Worsley (1986) and Siegmund (1988), while traditional Bayesian analysis and *Markov chain Monte Carlo* (MCMC) hierarchical Bayesian analysis are presented in Raftery and Akman (1986) and Carlin, Gelfand and Smith (1992), respectively. Arnold (1993) considered the application of the Gibbs sampler to a Poisson distribution with a changepoint. For binomial CPPs, Smith (1975) presented the conventional Bayesian approach for a finite sequence of independent observations with details on binomial single-changepoint model. Smith (1980) studied binomial multiple-changepoint model, which were investigated by Stephens (1994) using the Gibbs sampler. For binary CPPs, Halpern (1999) applied a novel changepoint statistic based on the minimum value, over possible changepoint locations of Fisher's Exact Test to assessing recombination in genetic sequences of HIV. For multiple change-point models, Chib (1998) provided a comparison study and Fearnhead and Liu (2007) proposed an on-line algorithm. Three comprehensive reviews on CPPs are provided by Brodsky and Darkhovsky (1993), Chen and Gupta (2000) and more recently by Wu (2005).

The primary objective in the analysis of CPPs is to make inferences about unknown changepoints and the associated parameters. Although the MCMC methods can be employed in such Bayesian analyses, in our viewpoint, the difficulties lie in monitoring the convergence of the Markov chains. In addition, they could suffer from slow convergence. These issues have prompted some researchers to take the view that the MCMC methods are to be used only when there is no better alternative (see, e.g., discussions in Evans and Swartz (1995, 2000) and Hobert and Casella (1996)). In this article, we first propose a new non-iterative Bayesian sampling approach (called exact IBF sampling), which completely avoids the problem of convergence and slow convergence. The idea is to first utilize the sampling-wise *inverse Bayes formulae* (IBF, Tan *et al*., 2003) to derive the conditional distribution of the missing data given the observed data, and then to draw iid samples from the complete-data posterior distribution.

In practice, we are generally uncertain about the number of changepoints. Hence, model determination is the first task in changepoint analysis. Let M_s represent a model with s changepoints. A classical approach of selecting the most appropriate model is the likelihood ratio test by comparing M_s with M_{s+1} (e.g., Henderson and Matthews, 1993). Gelfand and Dey (1994) reviewed the behavior of the likelihood ratio statistic and well-known adjustments to it. In the context of Bayesian analysis, Bayes factor is a useful tool for model choice. However, the calculation of Bayes factor itself has proved extremely challenging (Kass and Raftery, 1995). Approximate computation of Bayes factor (equivalently, marginal likelihood) can be implemented by using the Gibbs output (Chib, 1995) or the more general MCMC output (Chen,

2005). In this paper, we will we develop two alternative formulae to exactly calculate marginal likelihood (or Bayes factor) by using the exact IBF output and the point-wise IBF (Tan *et al*., 2003), respectively.

The rest of this paper is organized as follows. In Section 2, we formulate the general Bayesian changepoint models. In Section 3, we propose a non-iterative Bayesian sampling approach (called the exact IBF sampling) and derive two simple formulae to exactly calculate the marginal likelihood. Section 4 considers Poisson models with single and multiple changepoints and the corresponding Bayesian model selection. In Section 5, we re-analyze the HUS data using the proposed methods. Simulations are conducted to validate the performance of the proposed methods in Section 6. Conclusion and comment are presented in Section 7.

2. Bayesian formulation for changepoint problems

Let $Y_{\text{obs}}=\{y_i\}_{i=1}^n$ denote a realization of the sequence of independent random variables $\{Y_i\}_{i=1}^n$ of length n. The random variables $\{Y_i\}_{i=1}^n$ are said to have a changepoint at r ($1 \le r \le n$) if $Y_i \sim f_1(y|\theta_1)$ ($i = 1,\ldots, r$) and $Y_i \sim f_2(y|\theta_2)$ ($i = r + 1,\ldots, n$), where $f_1(y|\theta_1) \neq f_2(y|\theta_2)$, θ_1 and θ_2 could be vector-valued. In particular, the point $r = n$ represents 'no change'. Thus, the likelihood function becomes

$$L(Y_{\text{obs}}|r,\theta_1,\theta_2)=\Pi_{i=1}^{r} f_1(y_i|\theta_1)\cdot\Pi_{i=r+1}^{n} f_2(y_i|\theta_2). \tag{2.1}$$

Using $\pi(r, \theta_1, \theta_2)$ as a joint prior distribution for r, θ_1, and θ_2, the joint posterior distribution is given by

$$f(r,\theta_1,\theta_2|Y_{\text{obs}}) \propto L(Y_{\text{obs}}|r,\theta_1,\theta_2)\cdot\pi(r,\theta_1,\theta_2). \tag{2.2}$$

This single-changepoint problem can be easily generalized to incorporate multiple changes in the sequence. The Bayesian formulation for the multiple-changepoint problem is almost identical with that for the single-changepoint problem. Let M_s represent a model with s changepoints denoted by $\boldsymbol{r} = (r_1,\ldots, r_s)^\top$. Similar to (2.2), under M_s (s is given), we have

$$\begin{aligned} f(\boldsymbol{r},\theta_1,\ldots,\theta_{s+1}|Y_{\text{obs}}) &\propto L(Y_{\text{obs}}|\boldsymbol{r},\theta_1,\ldots,\theta_{s+1})\cdot\pi(\boldsymbol{r},\theta_1,\ldots,\theta_{s+1}) \\ &=\{\Pi_{j=1}^{s+1}\Pi_{i=r_{j-1}+1}^{r_j} f_j(y_i|\theta_j)\}\cdot\pi(\boldsymbol{r},\theta_1,\ldots,\theta_{s+1}), \end{aligned} \tag{2.3}$$

where $r_0 \equiv 0$, $r_{s+1} \equiv n$, and the changepoints $\boldsymbol{r}$ take values in the domain

$$\mathcal{S}(\boldsymbol{r}|Y_{\text{obs}})=\{\boldsymbol{r}: 1 \le r_1 < \cdots < r_s \le n, r_j \text{ is an integer}, j=1,\ldots,s\}. \tag{2.4}$$

The primary objective is to make inferences about the unknown changepoints $\boldsymbol{r}$ and the unknown parameters $(\theta_1,\ldots, \theta_{s+1})$.

3. Exact IBF sampling and marginal likelihood calculation

For a given model, let Y_{obs} denote the observed data, Z the missing data or latent data (e.g., changepoints) and θ the model-specific parameter vector. We further denote the likelihood

function by $L(Y_{obs}|\theta)$ and the marginal density of Y_{obs} (equivalently, the marginal likelihood) by $m(Y_{obs})$. Within a Bayesian framework, we assume that the prior density is $\pi(\theta)$. Two basic tasks are (i) for the purpose of Bayesian inferences, how to obtain iid samples from the observed posterior $f(\theta|Y_{obs})$ or equivalently from the joint posterior $f(Z, \theta|Y_{obs})$, and (ii) for the purpose of Bayesian model choice, how to exactly calculate the marginal likelihood $m(Y_{obs})$ for the given model.

3.1 Exact IBF sampling

In general, we can obtain explicit expressions for both the complete-data posterior distribution $f(\theta|Y_{obs}, Z)$ and the conditional predictive distribution $f(Z|Y_{obs}, \theta)$, that is, the sampling from the two distributions and the evaluation of the two densities can routinely be implemented. The fundamental conditional sampling principle implies

$$f(Z, \theta|Y_{obs}) = f(Z|Y_{obs}) \cdot f(\theta|Y_{obs}, Z),$$

which states that if we could draw $Z^{(\ell)} \overset{\text{iid}}{\sim} f(Z|Y_{obs})$ and simulate $\theta^{(\ell)} \sim f(\theta|Y_{obs}, Z^{(\ell)})$, then $\{(Z^{(\ell)}, \theta^{(\ell)})\}_1^L$ are iid samples from the joint posterior $f(Z, \theta|Y_{obs})$. Therefore, the key is to be able to generate iid samples from $f(Z|Y_{obs})$.

Let $\mathcal{S}(\theta|Y_{obs})$ and $\mathcal{S}(Z|Y_{obs})$ denote the conditional supports of $\theta|Y_{obs}$ and $Z|Y_{obs}$, respectively. The sampling IBF shows that (Tan *et al*., 2003)

$$f(Z|Y_{obs}) \propto \frac{f(Z|Y_{obs}, \theta_0)}{f(\theta_0|Y_{obs}, Z)}, \quad \begin{array}{l}\text{for an arbitrary } \theta_0 \in \mathcal{S}(\theta|Y_{obs}) \\ \text{and all } Z \in \mathcal{S}(Z|Y_{obs}).\end{array} \tag{3.1}$$

Consider the case where Z is a discrete random variable/vector taking finite values on the conditional support $\mathcal{S}(Z|Y_{obs})$. For example, in (2.2), the changepoint r takes values in $\{1,\ldots, n\}$; and in (2.3), the s changepoints $(r_1,\ldots, r_s)$ take values in $\mathcal{S}(\boldsymbol{r}|Y_{obs})$ defined by (2.4). Without loss of generality, we denote the conditional support of $Z|(Y_{obs}, \theta)$ by $\mathcal{S}(Z|Y_{obs}, \theta) = \{z_1,\ldots, z_K\}$. Since $f(Z|Y_{obs}, \theta)$ is available, firstly, we can directly identify $\{z_k\}_1^K$ from the model specification and thus all $\{z_k\}_1^K$ are known. Secondly, we assume that $\{z_k\}_1^K$ do not depend on the value of θ, therefore, we have

$$\mathcal{S}(Z|Y_{obs}) = \mathcal{S}(Z|Y_{obs}, \theta) = \{z_1, \ldots, z_K\}.$$

Because of the discreteness of Z, the notation $f(z_k|Y_{obs})$ will used to denote the pmf, i.e., $f(z_k|Y_{obs}) = \Pr\{Z = z_k|Y_{obs}\}$. Thus, the key is to find $p_k = f(z_k|Y_{obs})$ for $k = 1, \ldots, K$. For some $\theta_0 \in \mathcal{S}(\theta|Y_{obs})$, let

$$q_k = q_k(\theta_0) = \Pr\{Z = z_k|Y_{obs}, \theta_0\} / f(\theta_0|Y_{obs}, z_k), \quad k = 1, \ldots, K. \tag{3.2}$$

As both $f(Z|Y_{obs}, \theta)$ and $f(\theta|Y_{obs}, Z)$ are available, the computation of (3.2) is straight-forward. Observing that all $\{q_k\}_1^K$ depend on θ_0, we denote q_k by $q_k(\theta_0)$ to emphasize its dependency on θ_0. From the sampling IBF (3.1), we obtain

$$p_k = q_k(\theta_0) / \sum_{k'=1}^{K} q_{k'}(\theta_0), \quad k = 1, \ldots, K, \tag{3.3}$$

where $\{p_k\}_1^K$ do not depend on θ_0 since $\{p_k\}_1^K$ are normalizing probabilities of $\{q_k\}_1^K$. Thus, it is easy to sample from $f(Z|Y_{obs})$ since it is a discrete distribution with probabilities $\{p_k\}_1^K$ on $\{z_k\}_1^K$ (e.g., the built-in S-plus function "`sample`" is especially designed for this purpose). We summarize the algorithm as follows.

The exact ibf sampling—Given both the complete-data posterior distribution $f(\theta|Y_{obs}, Z)$ and the conditional predictive distribution $f(Z|Y_{obs}, \theta)$,

- **a.** Identify $\mathcal{S}(Z|Y_{obs}) = \{z_1, \ldots, z_K\}$ from $f(Z|Y_{obs}, \theta)$ and calculate $\{p_k\}_1^K$ according to (3.3) and (3.2);
- **b.** Generate iid samples $\{Z^{(\ell)}\}_{\ell=1}^L$ of Z from the pmf $f(Z|Y_{obs})$ with probabilities $\{p_k\}_1^K$ on $\{z_k\}_1^K$;
- **c.** Generate $\theta^{(\ell)} \sim f(\theta|Y_{obs}, Z^{(\ell)})$ for $\ell = 1, \ldots, L$, then $\{\theta^{(\ell)}\}_1^L$ are iid samples from the observed posterior distribution $f(\theta|Y_{obs})$.

3.2 Exact calculation of marginal likelihood

In this subsection, we provide two alternative formulae to calculate the marginal likelihood $m(Y_{obs})$. Let $\{(Z^{(\ell)}, \theta^{(\ell)})\}_1^L$ denote the output from the exact IBF sampling. From Bayes formula: $m(Y_{obs}) = L(Y_{obs}|\theta)\pi(\theta)/f(\theta|Y_{obs})$, which holds for any θ, we have

$$\log m(Y_{obs}) = \log L(Y_{obs}|\theta_0) + \log \pi(\theta_0) - \log f(\theta_0|Y_{obs}), \quad \theta_0 \in \mathcal{S}(\theta|Y_{obs}). \tag{3.4}$$

For estimation efficiency, θ_0 is generally taken to be a high-density point in the support of the posterior (e.g, the posterior mode/mean as suggested by Chib (1995)). Since the observed posterior density can be written as

$$f(\theta|Y_{obs}) = \int f(\theta|Y_{obs}, Z) f(Z|Y_{obs}) dZ,$$

we obtain a Monte Carlo estimate of $f(\theta|Y_{obs})$ at θ_0,

$$\widehat{f}(\theta_0|Y_{obs}) = (1/L) \sum_{\ell=1}^{L} f(\theta_0|Y_{obs}, Z^{(\ell)}), \tag{3.5}$$

where $\{Z^{(\ell)}\}$ are iid samples from $f(Z|Y_{obs})$. Note that this estimate is simulation consistent, i.e., $\hat{f}(\theta_0|Y_{obs}) \to f(\theta_0|Y_{obs})$ as $L \to \infty$. Combining (3.4) with (3.5), we have an approximate formula to calculate $m(Y_{obs})$.

On the other hand, note that Z is a discrete random variable taking values on $\{z_k\}_1^K$, using the point-wise IBF (Tan *et al*., 2003): $f(\theta|Y_{obs}) = \{\int f(Z|Y_{obs}, \theta)/f(\theta|Y_{obs}, Z)\, dZ\}^{-1}$, we explicitly have

$$f(\theta_0|Y_{\text{obs}})=\{\sum_{k'=1}^{K} q_{k'}(\theta_0)\}^{-1}=p_1/q_1(\theta_0), \tag{3.6}$$

where p_k and $q_k(\theta_0)$ are defined in (3.3) and (3.2), respectively. Substituting (3.6) into (3.4) gives an exact formula to calculate $m(Y_{\text{obs}})$.

4. Poisson models with changepoints

4.1 A single changepoint

Let M_s represent a model with s changepoints, Poisson(θ) a Poisson distribution with mean θ and Poisson($\cdot|\theta$) the corresponding probability mass function. We first consider the single-changepoint model M_1. In (2.1), we let $f_j(y|\theta_j)$ = Poisson($y|\theta_j$) for j = 1, 2. As a joint prior distribution for (r, θ_1, θ_2), we assume that r, θ_1 and θ_2 are independent, r has a discrete uniform distribution on {1,…,n},

$$\theta_1 \sim \text{Ga}(a_1, b_1) \quad \text{and} \quad \theta_2 \sim \text{Ga}(a_2, b_2), \tag{4.1}$$

where Ga(a, b) is a gamma distribution with density $\text{Ga}(x|a, b) = b^a x^{a-1} e^{-bx}/\Gamma(a)$, $x \geq 0$. Thus, the joint posterior distribution (2.2) becomes

$$f(r, \theta_1, \theta_2|Y_{\text{obs}}) \propto \theta_1^{a_1+S_r-1} e^{-(b_1+r)\theta_1} \cdot \theta_2^{a_2+S_n-S_r-1} e^{-(b_2+n-r)\theta_2},$$

where $S_r \equiv \sum_{i=1}^{r} y_i$. Direct calculation yields

$$f(\theta_1, \theta_2|Y_{\text{obs}}, r)=\text{Ga}(\theta_1|a_1+S_r,\ b_1+r) \cdot \text{Ga}(\theta_2|a_2+S_n-S_r,\ b_2+n-r), \tag{4.2}$$

$$f(r|Y_{\text{obs}}, \theta_1, \theta_2)=\frac{(\theta_1/\theta_2)^{S_r}\exp\{(\theta_2-\theta_1)r\}}{\sum_{i=1}^{n}(\theta_1/\theta_2)^{S_i}\exp\{(\theta_2-\theta_1)i\}}, \quad r=1,\ldots,n. \tag{4.3}$$

We can treat the changepoint r as latent variable Z and (θ_1, θ_2) as parameter vector θ. By using (3.1)–(3.3), for any given ($\theta_{1,0}$, $\theta_{2,0}$) ∈ $\mathcal{S}(\theta_1, \theta_2|Y_{\text{obs}})$, we immediately obtain

$$f(r|Y_{\text{obs}})=\frac{\Gamma(a_1+S_r)\Gamma(a_2+S_n-S_r)/[(b_1+r)^{a_1+S_r}(b_2+n-r)^{a_2+S_n-S_r}]}{\sum_{i=1}^{n}\Gamma(a_1+S_i)\Gamma(a_2+S_n-S_i)/[(b_1+i)^{a_1+S_i}(b_2+n-i)^{a_2+S_n-S_i}]}, \tag{4.4}$$

where r = 1,…,n. It again confirms that the right-hand side of (4.4) does not depend on ($\theta_{1,0}$, $\theta_{2,0}$). Based on (4.4) and (4.2), we can obtain iid posterior samples for the changepoint r and the parameters (θ_1, θ_2) by using the exact IBF sampling.

4.2 Multiple changepoints

Now we consider the multiple-changepoints model M_s. In (2.3), let $f_j(y|\theta_j) = \text{Poisson}(\theta_j)$ for $j = 1,\ldots, s+1$, where $\theta = (\theta_1,\ldots, \theta_{s+1})^\top$ is the mean vector and $\boldsymbol{r} = (r_1,\ldots, r_s)^\top$ denote the s changepoints taking integer values on the domain $\mathcal{S}(\boldsymbol{r}|Y_{\text{obs}})$ defined in (2.4). We use independent priors for $\boldsymbol{r}$, θ, and $\boldsymbol{r}$ has a discrete uniform prior on $\mathcal{S}(\boldsymbol{r}|Y_{\text{obs}})$,

$$\theta_j \sim \text{Ga}(a_j, b_j), \quad j=1,\ldots, s+1. \tag{4.5}$$

Thus, the joint posterior distribution (2.3) becomes

$$f(\boldsymbol{r}, \theta|Y_{\text{obs}}) \propto \Pi_{j=1}^{s+1}\theta_j^{a_j+S_{r_j}-S_{r_{j-1}}-1} e^{-(b_j+r_j-r_{j-1})\theta_j}, \tag{4.6}$$

where $S_r \equiv \sum_{i=1}^{r} y_i$, $r_0 \equiv 0$ and $r_{s+1} \equiv n$. From (4.6), we have

$$f(\theta|Y_{\text{obs}}, \boldsymbol{r})=\Pi_{j=1}^{s+1}\text{Ga}(\theta_j|a_j+S_{r_j} - S_{r_{j-1}}, b_j+r_j - r_{j-1}), \tag{4.7}$$

$$f(\boldsymbol{r}|Y_{\text{obs}}, \theta) \propto \Pi_{j=1}^{s}(\theta_j/\theta_{j+1})^{S_{r_j}} e^{(\theta_{j+1}-\theta_j)r_j}, \quad \boldsymbol{r} \in \mathcal{S}(\boldsymbol{r}|Y_{\text{obs}}). \tag{4.8}$$

We treat $\boldsymbol{r}$ as latent variables and θ as parameter vector. By using (3.1)–(3.3), for any given $\theta_0 \in \mathcal{S}(\theta|Y_{\text{obs}})$, we immediately obtain

$$f(\boldsymbol{r}|Y_{\text{obs}}) \propto \prod_{j=1}^{s+1} \frac{\Gamma(a_j+S_{r_j} - S_{r_{j-1}})}{(b_j+r_j - r_{j-1})^{a_j+S_{r_j}-S_{r_{j-1}}}}, \quad \boldsymbol{r} \in \mathcal{S}(\boldsymbol{r}|Y_{\text{obs}}). \tag{4.9}$$

Based on (4.9) and (4.7), we can obtain iid posterior samples for the changepoints $\boldsymbol{r}$ and the parameter vector θ by using the exact IBF sampling.

4.3 Determining the number of changepoints via Bayes factor

In practice, we are generally uncertain about the number of changepoints. Hence, model determination is the first task in changepoint analysis. Let M_s represent the Poisson model with s changepoints $\boldsymbol{r} = (r_1,\ldots, r_s)^\top$ and $\theta = (\theta_1,\ldots, \theta_{s+1})^\top$ the mean vector. Further let $\Theta = (\boldsymbol{r}, \theta)$ and $\hat{\Theta} = (\hat{\boldsymbol{r}}, \hat{\theta})$ denote the posterior means obtained via the exact IBF output. Under model M_s, from (3.4), the log-marginal likelihood is given by

$$\log m(Y_{\text{obs}}|M_s)=\log L(Y_{\text{obs}}|\widehat{\Theta}, M_s)+\log \pi(\widehat{\Theta}|M_s) - \log f(\widehat{\Theta}|Y_{\text{obs}}, M_s), \tag{4.10}$$

where $f(\hat{\Theta}|Y_{\text{obs}},M_s) = f(\hat{\boldsymbol{r}}|Y_{\text{obs}},M_s) \cdot f(\hat{\theta}|Y_{\text{obs}}, \hat{\boldsymbol{r}},M_s)$. We choose the model with the largest log-marginal likelihood. Essentially, the marginal likelihood approach is the same as the Bayes factor approach. A Bayes factor is defined as the ratio of posterior odds versus prior odds,

which is simply a ratio of two marginal likelihoods. For comparing models M_s and M_{s+1}, the Bayes factor for model M_s vs. model M_{s+1} is

$$B_{s,s+1}=\frac{m(Y_{\text{obs}}|M_s)}{m(Y_{\text{obs}}|M_{s+1})}. \tag{4.11}$$

Jeffreys (1961, Appendix B) suggested interpreting $B_{s,s+1}$ in half-units on the $\log_{10}$ scale, i.e., when $B_{s,s+1}$ falls in (1, 3.2), (3.2, 10), (10, 100) and (100, $+\infty$), the evidence against M_{s+1} is not worth more than a bare mention, substantial, strong and decisive, respectively.

5. Analysis of the HUS data

We first analyze the Birmingham data in Table 1. Denote the number of cases of HUS in Birmingham in year i by y_i ($i = 1,\ldots,n$ with $n = 20$, and $i = 1$ denotes the year 1970). To determine the number of changepoints via Bayes factor, we can not use non-informative prior distributions because they are improper. We investigate models M_0, M_1 and M_2 and choose standard exponential prior distributions, specified by setting all $a_j = b_j = 1$ in (4.5). Based on (4.10), we calculate log-marginal likelihoods for the three models, and we obtain $\log m(Y_{\text{obs}}|M_0) = -86.14$, $\log m(Y_{\text{obs}}|M_1) = -57.56$ and $\log m(Y_{\text{obs}}|M_2) = -57.00$. Therefore, M_2 seems to be the most appropriate choice. From (4.11), the Bayes factor for M_1 versus M_0 is 2.583×10^{12}, while the Bayes factor for M_2 versus M_1 is 1.751. That is, the difference between M_2 and M_1 is not worth to mention. Therefore, we select M_1, which is consistent with the pattern indicated in Figure 1.

Under M_1, we assume that $y_1,\ldots,y_r \overset{\text{iid}}{\sim} \text{Poisson}(\theta_1)$ and $y_{r+1},\ldots,y_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta_2)$, where r is the unknown changepoint and $\theta_1 \neq \theta_2$. Table 2 contains the exact posterior probabilities for the changepoint r using (4.4). The changepoint occurs at $r = 11$ (i.e., year 1980) with posterior probability 0.9795. Based on (4.4) and (4.2), we generate 20, 000 iid posterior samples by using the exact IBF sampling, and the Bayes estimates of r, θ_1 and θ_2 are given by 11.013, 1.593 and 9.609. The corresponding Bayes standard errors are 0.143, 0.370 and 0.985. The 95% Bayes credible intervals for θ_1 and θ_2 are [0.952, 2.393] and [7.800, 11.621], respectively. Figures 2(a) and 2(b) show the histogram of the changepoint r and the posterior densities of θ_1 and θ_2, which are estimated by a kernel density smoother based on iid posterior samples. Figure 2(c) depicts the annual numbers of HUS for Birmingham series, the identified changepoint position, and the average number of cases before and after the changepoint.

Now we analyze the Newcastle data in Table 1. Similarly, three log-marginal likelihoods are given by $\log m(Y_{\text{obs}}|M_0) = -85.24$, $\log m(Y_{\text{obs}}|M_1) = -64.13$ and $\log m(Y_{\text{obs}}|M_2) = -64.10$. From (4.11), the Bayes factor for M_2 versus M_0 is 1.5169×10^9, and the Bayes factor for M_2 versus M_1 is 1.03. Therefore, we select M_2, which is consistent with the pattern as indicated in Figure 1. In addition, the selection of M_2 is also identical to that obtained by Henderson and Matthews (1993).

Under M_2, we assume that $y_1,\ldots,y_{r_1} \overset{\text{iid}}{\sim} \text{Poisson}(\theta_1)$, $y_{r_1+1},\ldots,y_{r_2} \overset{\text{iid}}{\sim} \text{Poisson}(\theta_2)$, and $y_{r_2+1},\ldots,y_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta_3)$, where (r_1, r_2) are the unknown changepoints and $\theta_1 \neq \theta_2 \neq \theta_3$. Using the standard exponential prior distributions, specified by letting $a_j = b_j = 1$ ($j = 1, 2, 3$) in (4.5), we obtained exact joint posterior probabilities for the changepoint pair (r_1, r_2) from (4.9). Two changepoints occur at $r_1 = 7$ and $r_2 = 15$ (i.e., year 1976 and year 1984) with the joint posterior probability being 0.3589. Based on (4.9) and (4.7), we generated 20, 000 iid posterior samples. The resulting Bayes estimates of r_1, r_2, θ_1, θ_2 and θ_3 are given by 7.638, 15.47, 1.805, 3.591 and 9.643. The 95% Bayes credible intervals for θ_1, θ_2 and θ_3 are [0.7461,

3.620], [1.5085, 11.50] and [0.2806, 13.32], respectively. Figures 3(a) and 3(b) display the histograms of r_1 and r_2. Figures 3(c) shows the posterior densities of θ_j (j = 1, 2, 3). Figure 3 (d) depicts the annual numbers of HUS, two identified changepoints, and the average number of cases before and after the two changepoints.

6. Simulation Studies

The first simulated dataset consists of 100 observations with $y_1, \ldots, y_{50} \overset{\text{iid}}{\sim} \text{Poisson}(3)$ and $y_{51}, \ldots, y_{100} \overset{\text{iid}}{\sim} \text{Poisson}(0.5)$. The simulated observations are shown in Figure 4(c). We again use standard exponential distributions as priors of θ. From (4.10), log-marginal likelihoods for three models M_0, M_1 and M_2 are given by $\log m(Y_{\text{obs}}|M_0) = -187.1$, $\log m(Y_{\text{obs}}|M_1) = -148.8$ and $\log m(Y_{\text{obs}}|M_2) = -149.3$. From (4.11), we have $B_{10} = 4.3 \times 10^{16}$ and $B_{12} = 1.649$, which suggest that M_1 is appropriate. Computations according to (4.4) show that the changepoint occurs at $r = 50$ with posterior probability 0.807. Based on (4.4) and (4.2), we generate 30, 000 iid posterior samples by using the exact IBF sampling. The Bayes means, standard errors, and 95% credible intervals for (r, θ_1, θ_2) are given by (50.6, 2.8226, 0.5249), (1.642, 0.240, 0.103) and [50, 56], [2.373, 3.315], [0.341, 0.742], respectively. Figures 4(a) and 4(b) show the histogram of r and the posterior densities of θ_1 and θ_2. Figure 4(c) displays the 100 simulated observations, the identified changepoint position, and the Bayes estimates of θ_1 and θ_2.

The second simulated dataset consists of 100 observations: $y_1, \ldots, y_{20} \overset{\text{iid}}{\sim} \text{Poisson}(5.5)$, $y_{21}, \ldots, y_{70} \overset{\text{iid}}{\sim} \text{Poisson}(0.8)$, and $y_{71}, \ldots, y_{100} \overset{\text{iid}}{\sim} \text{Poisson}(3.5)$. The simulated observations are shown in Figure 5(d). Similarly, we have $\log m(Y_{\text{obs}}|M_0) = -249.7$, $\log m(Y_{\text{obs}}|M_1) = -222.2$ and $\log m(Y_{\text{obs}}|M_2) = -185.6$, $B_{20} = 6.891 \times 10^{27}$, and $B_{21} = 7.856 \times$, which suggest that M_2 is appropriate. Computations according to (4.9) show that the changepoints occur at $r_1 = 20$ and $r_2 = 70$ with the joint posterior probability 0.7912. Based on (4.9) and (4.7), we generate 30, 000 iid posterior samples by using the exact IBF sampling. The Bayes means, standard errors, and 95% credible intervals for $(r_1, r_2, \theta_1, \theta_2, \theta_3)$ are given by (20.0091, 69.7871, 5.7120, 0.8427, 3.8190), (0.1019, 0.4457, 0.5230, 0.1296, 0.3517) and [20, 20], [69, 70], [4.735, 6.789], [0.610, 1.116], [3.161, 4.537], respectively. Figures 5(a) and 4(b) show the histogram of r_1 and r_2. Figures 5(c) shows the posterior densities of θ_1, θ_2 and θ_3. Figure 5(d) displays the 100 simulated observations, the two identified changepoint positions, and the Bayes estimates of θ_1, θ_2 and θ_3.

7. Discussion

It is noted that Barry and Hartigan (1992, 1993) and Fearnhead (2006) describe methods for calculating marginal likelihoods for multiple change-point problems. The latter also discusses methods for simulating from the change-point positions. Barry and Hartigan (1992, 1993) assumed a specific prior structure on the number and position of changepoints; while Fearnhead (2006) considered models with a fixed number of change-points. Although it was claimed that these methods can deal with arbitrarily large numbers of change-points, these methods are quite complicated in implementation. For example, the parameter values may be estimated exactly in $O(n^3)$ calculations, or to an adequate approximation by MCMC methods that are $O(n)$ in the number of observations.

In this paper, we considered Poisson changepoint analysis by using an exact IBF sampling approach. The advantages of the proposed exact IBF sampling method over MCMC methods are that: (i) there is no requirement to diagnose whether the MCMC algorithms has converged, i.e., the former entirely avoids the problems of convergence and slow convergence associated with MCMC methods; (ii) because the samples generated from the observed posterior distribution are independent it is straightforward to quantify uncertainty in estimates of features

of the posterior distributions based on them. To determine the number of changepoints, we developed simple methods to exactly calculate marginal likelihood (or Bayes factor). Two simulations are conducted to validate the performance of the proposed methods.

We should point out that the proposed approach is limited. For example, let M_s represent a model with s changepoints and the s-changepoint is denoted by $\boldsymbol{r} = (r_1,\ldots, r_s)^{\top}$. For a large number of observations or $s \geq 4$, the calculation of (4.9) becomes prohibitive. In this cases, the general IBF sampler (Tian *et al*., 2003) is a feasible alternative.

In the re-analysis of the HUS data using the proposed methods, we have focused on the annual numbers of cases rather than the incidence of the HUS because accurate population are difficult to obtain for the catchment areas. Possibilities for further analysis might consider extra-Poisson variation, treads in mean, the influence of some covariates (e.g., age, race) and so on.

Acknowledgments

We are grateful to the Editor, an Associate Editor and three referees for their constructive comments and suggestions. GL Tian and M Tan's research was supported in part by U.S. National Cancer Institute grants CA106767 and CA119758. The research of KW Ng was partially supported by a research grant of the University of Hong Kong. Special thanks should go to one referee for drawing our attention to several recent papers on multiple change-point problems.

References

Arnold, SF. Gibbs sampler. In: Rao, CR., editor. Handbook of Statistics. Vol. 9. Elsevier Science Publishers B. V; 1993. p. 599-625.

Barry D, Hartigan JA. Product partition models for change point problems. Ann Statist 1992;20:260–279.

Barry D, Hartigan JA. A Bayesian analysis for change point problems. Journal of the American Statistical Association 1993;88:309–319.

Brodsky, BE.; Darkhovsky, BS. Series: Mathematics and Its Applications. Vol. 243. Springer; New York: 1993. Nonparametric Methods in Change Point Problems.

Carlin BP, Gelfand AE, Smith AFM. Hierarchical Bayesian analysis of changepoint problems. Appl Statist 1992;41:389–405.

Chen MH. Computing marginal likelihoods from a single MCMC output. Statistica Neerlandica 2005;59:16–29.

Chen, J.; Gupta, AK. Parametric Statistical Change Point Analysis. Springer; New York: 2000.

Chib S. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 1995;90:1313–1321.

Chib S. Estimation and comparison of multiple change-point models. Journal of Econometrics 1998;86:221–241.

Evans M, Swartz T. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems (with discussions). Statist Sci 1995;10:254–272.

Evans, M.; Swartz, T. Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press; Oxford: 2000.

Fearnhead P. Exact and efficient inference for multiple changepoint problems. Statist Comput 2006;16:203–213.

Fearnhead P, Liu Z. On-line inference for multiple changepoint problems. Journal of the Royal Statistical Society, Series B 2007;64:589–605.

Gelfand AE, Dey DK. Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society, Series B 1994;56:501–514.

Halpern AL. Minimally selected *p* and other tests for a single abrupt changepoint in a binary sequence. Biometrics 1999;55:1044–1050. [PubMed: 11315046]

Henderson R, Matthews JNS. An investigation of changepoints in the annual number of cases of haemolytic uraemic syndrome. Appl Statist 1993;42:461– 471.

Hobert JP, Casella G. The effect of improper priors on Gibbs sampling in hierarchical linear models. J Am Statist Assoc 1996;91:1461–1473.

Jarrett RG. A note on the intervals between coal-mining disasters. Biometrika 1979;66:191–193.

Jeffreys, H. Theory of Probability. 3. Oxford: Oxford University Press; 1961.

Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association 1995;90:773–795.

Maguire BA, Pearson ES, Wynn AHA. The time intervals between industrial accidents. Biometrika 1952;38:168–180.

Milford DV, Taylor CM, Gutteridge B, Hall SM, Rowe B, Kleanthous H. Haemolytic uraemic syndromes in the British Isles 1985–8: association with verocytotoxin producing Escherichia coli. Part 1: Clinical and epidemiological aspects. Arch Dis Child 1990;65(7):716–721. [PubMed: 2201261]

Raftery AE, Akman VE. Bayesian analysis of a Poisson process with a change-point. Biometrika 1986;73:85–89.

Siegmund D. Confidence sets in change point problems. Int Statist Rev 1988;56:31–48.

Smith AFM. A Bayesian approach to inference about a change-point in a sequence of random variables. Biometrika 1975;62:407–416.

Smith, AFM. Change-point problems: Approaches and applications. In: Bernardo, JM.; DeGroot, MH.; Lindley, DV.; Smith, AFM., editors. Bayesian Statistics. Vol. 1. Valencia: Valencia University Press; 1980. p. 83-89.

Smith AFM, Cook DG. Straight lines with a change-point: A Bayesian analysis of some renal transplant data. Appl Statist 1980;29:180–189.

Stephens DA. Bayesian retrospective multiple-changepoint identification. Appl Statist 1994;43:159–178.

Tan M, Tian GL, Ng KW. A noniterative sampling method for computing posteriors in the structure of EM-type algorithms. Statistica Sinica 2003;13:625–639.

Tarr PI, Neill MA, Allen J, Siccardi CJ, Watkins SL, Hickman RO. The increasing incidence of the hemolytic-uremic syndrome in King County, Washington: lack of evidence for ascertainment bias. Am J Epidemiol 1989;129(3):582–586. [PubMed: 2916551]

Worsley KJ. Confidence regions and tests for a change-point in a sequence of exponential family random variables. Biometrika 1986;73:91–104.

Wu, YH. Series: Lecture Notes in Statistics. Vol. 180. Springer; New York: 2005. Inference for Change Point and Post Change Means After a CUSUM Test.

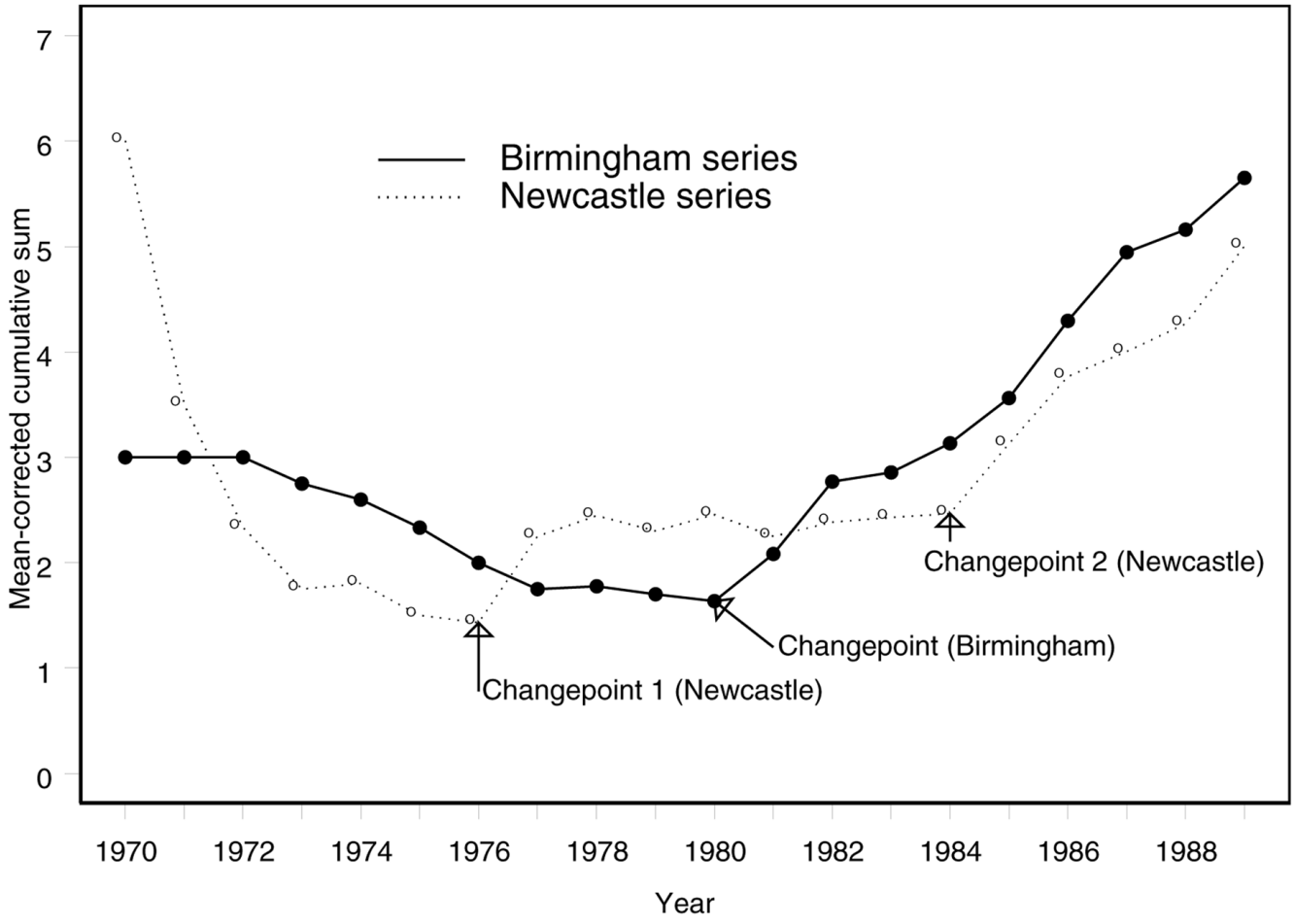

Figure 1.
Mean-corrected cumulative sum plot for the number of cases at Birmingham and Newcastle.

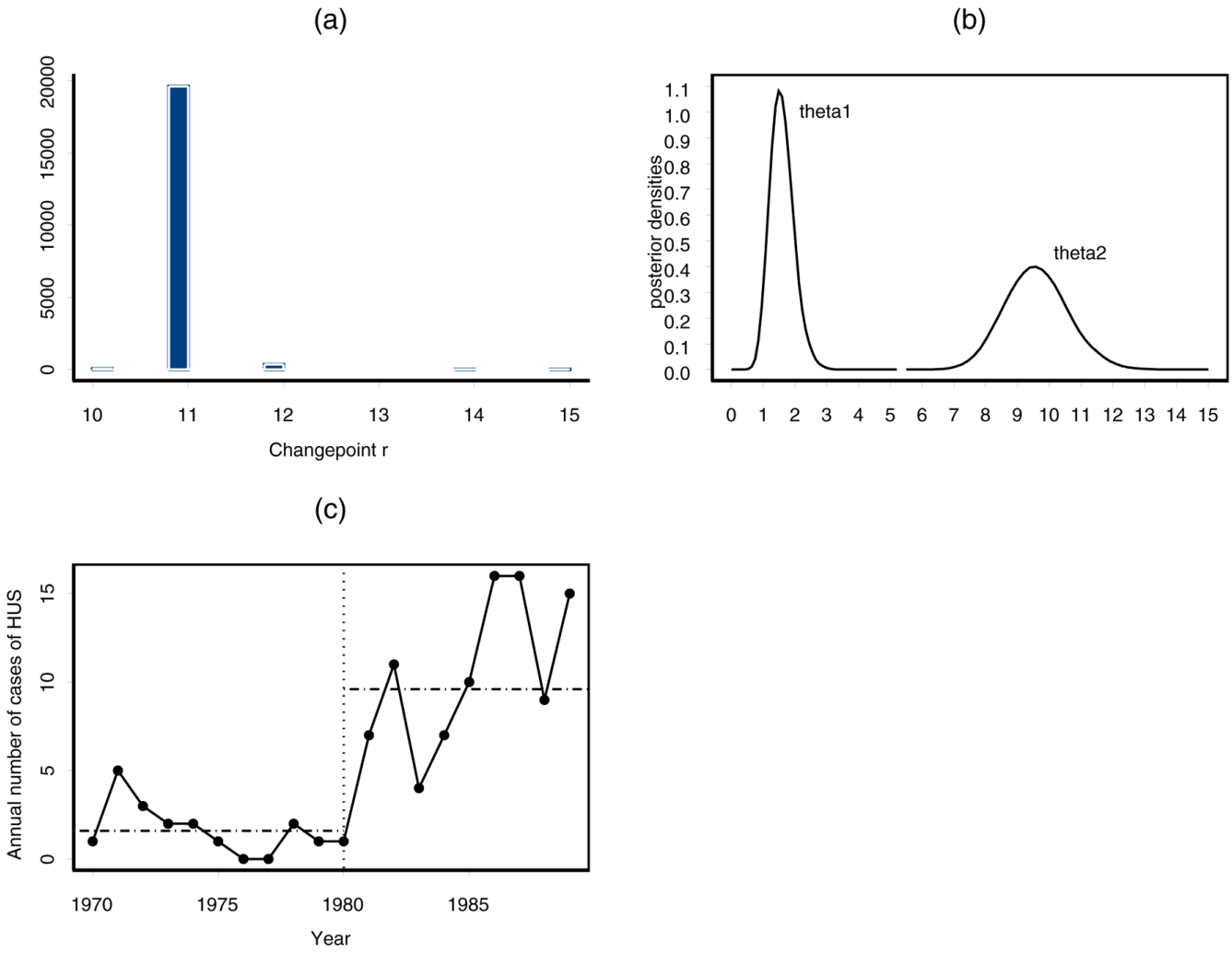

Figure 2.

Birmingham data set. (a) Histogram of the changepoint r. (b) The posterior densities of θ_1 and θ_2 estimated by a kernel density smoother based on 20, 000 iid samples generated via the exact IBF sampling. (c) The annual numbers of cases of HUS from 1970 to 1989. The dotted vertical line denotes the identified changepoint position, the lower horizontal line the average number (1.593) of cases during 1970–1980, and the upper horizontal line the average number (9.609) of cases during 1980–1989.

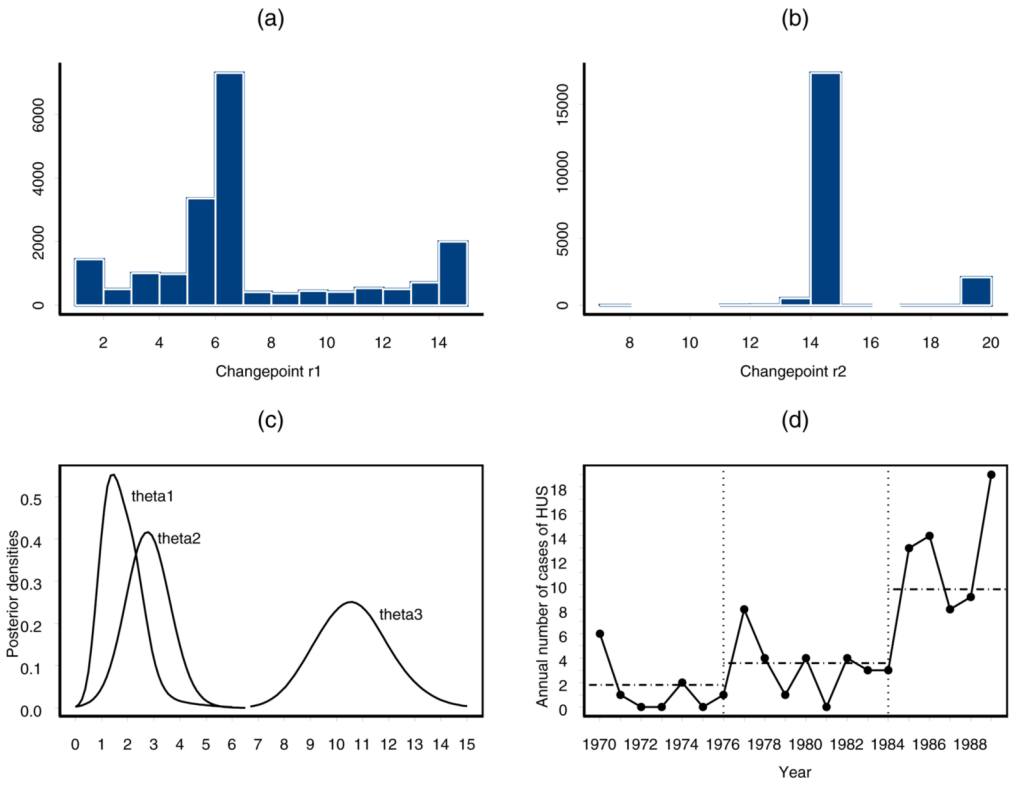

Figure 3.

Newcastle data set. (a) Histogram of the changepoint r_1. (b) Histogram of the changepoint r_2. (c) The posterior densities of θ_1, θ_2 and θ_3 estimated by a kernel density smoother based on 20, 000 iid samples generated via the exact IBF sampling. (d) The annual numbers of cases of HUS at Newcastle from 1970 to 1989. The two vertical lines denote two identified changepoint positions (1976 (1984), the three horizontal lines the average numbers (1.805, 3.591, 9.643) of cases during 1970–1976, 1976–1984 and 1984–1989, respectively.

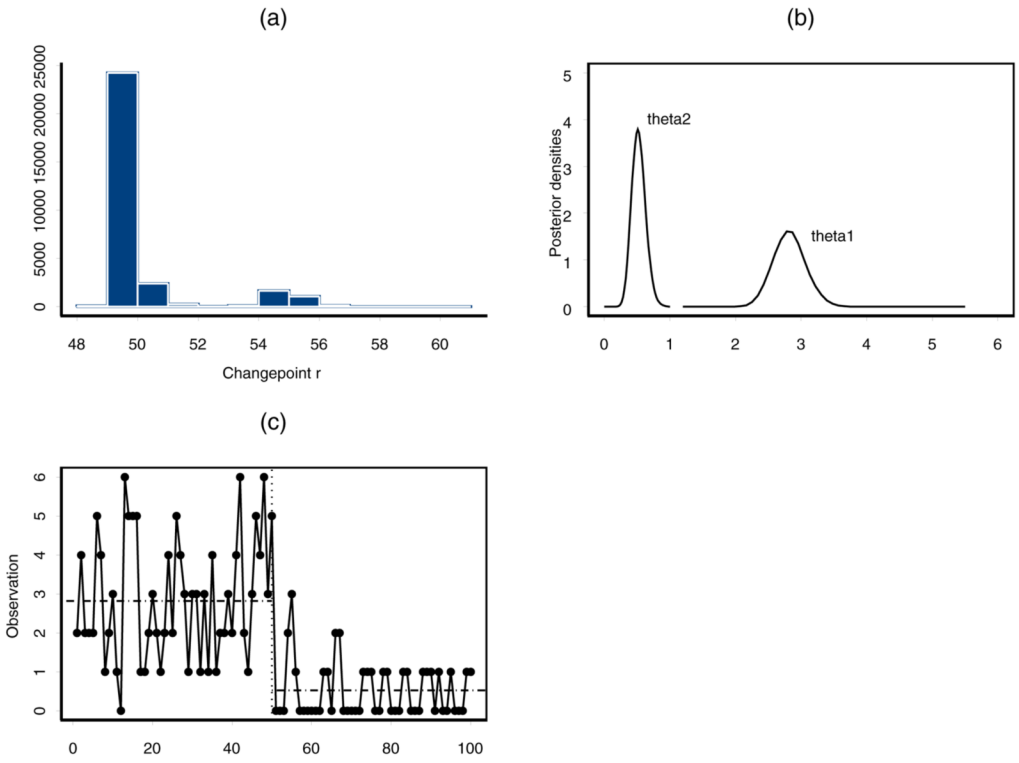

Figure 4.

Simulated dataset with one changepoint. (a) Histogram of r. (b) The posterior densities of θ_1 and θ_2. (c) The 100 simulated observations. The dotted vertical line denotes the identified changepoint position ($r = 50$), the left horizontal line the Bayes estimate of θ_1 ($\hat{\theta}_1 = 2.8226$), and the right horizontal line the Bayes estimate of $\hat{\theta}_2$ ($\hat{\theta}_2 = 0.5249$).

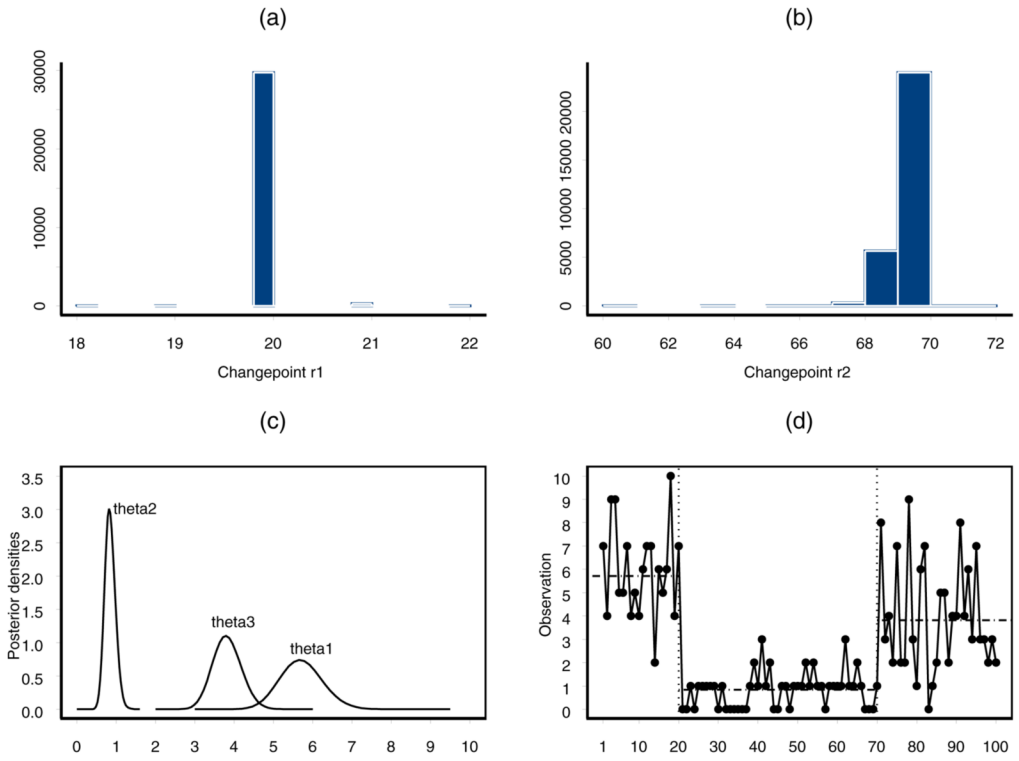

Figure 5.
Simulated dataset with two changepoints. (a) Histogram of r_1. (b) Histogram of r_2. (c) The posterior densities of θ_1, θ_2 and θ_3. (d) The 100 simulated observations, two identified changepoints (20, 70), and three Bayes estimates ($\hat{\theta}_1 = 5.7120$, $\hat{\theta}_2 = 0.8427$ and $\hat{\theta}_3 = 3.8190$).

Table 1

Counts of cases of HUS at Birmingham and Newcastle (Tarr *et al.*, 1989)

Observation	Year	Count at Birmingham	Count at Newcastle	Observation	Year	Count at Birmingham	Count at Newcastle
1	1970	1	6	11	1980	1	4
2	1971	5	1	12	1981	7	0
3	1972	3	0	13	1982	11	4
4	1973	2	0	14	1983	4	3
5	1974	2	2	15	1984	7	3
6	1975	1	0	16	1985	10	13
7	1976	0	1	17	1986	16	14
8	1977	0	8	18	1987	16	8
9	1978	2	4	19	1988	9	9
10	1979	1	1	20	1989	15	19

Table 2

Exact posterior probabilities for the changepoint r for Birmingham series

r	1	2	3	4	5
$f(r\|Y_{obs})$	2.249×10^{-13}	1.499×10^{-14}	2.651×10^{-14}	1.365×10^{-13}	8.493×10^{-13}
r	6	7	8	9	10
$f(r\|Y_{obs})$	1.994×10^{-11}	2.669×10^{-09}	7.541×10^{-07}	1.656×10^{-05}	2.899×10^{-03}
r	11	12	13	14	15
$f(r\|Y_{obs})$	9.795×10^{-01}	1.753×10^{-02}	3.628×10^{-06}	3.020×10^{-05}	7.756×10^{-06}
r	16	17	18	19	20
$f(r\|Y_{obs})$	8.459×10^{-08}	4.596×10^{-12}	4.673×10^{-15}	1.404×10^{-15}	1.952×10^{-14}