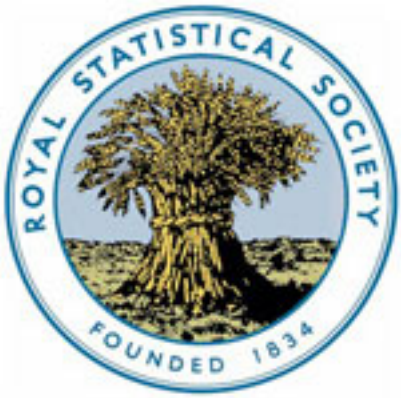

Appl. Statist. (1992)
41, *No.* 2, *pp.* 389–405

Hierarchical Bayesian Analysis of Changepoint Problems

By BRADLEY P. CARLIN,

University of Minnesota, Minneapolis, USA

ALAN E. GELFAND†

University of Connecticut, Storrs, USA

and ADRIAN F. M. SMITH

Imperial College of Science, Technology and Medicine, London, UK

[Received March 1990. Revised April 1991]

SUMMARY

A general approach to hierarchical Bayes changepoint models is presented. In particular, desired marginal posterior densities are obtained utilizing the Gibbs sampler, an iterative Monte Carlo method. This approach avoids sophisticated analytic and numerical high dimensional integration procedures. We include an application to changing regressions, changing Poisson processes and changing Markov chains. Within these contexts we handle several previously inaccessible problems.

Keywords: Changepoints; Changing Markov chains; Changing Poisson processes; Changing regressions; Gibbs sampler; Hierarchical Bayes models

1. Introduction

The literature on changepoint problems is, by now, enormous. Here we consider only the so-called non-sequential or fixed sample size version although an informal sequential procedure which follows from Smith (1975) is a routine consequence (see Section 6). Still the literature is substantial and we merely note several reviews which span both parametric and nonparametric approaches. These are Hinkley *et al.* (1980), Siegmund (1986), Wolfe and Schechtman (1984) and Zacks (1983).

Our focus is on a fully Bayesian parametric approach. Use of the Bayesian framework for inference with regard to the changepoint dates to work by Chernoff and Zacks (1964) and Shiryayev (1963). Smith (1975) presents the Bayesian formulation for a finite sequence of independent observations. In particular he addresses three situations:

(a) both the initial distribution and the changed distribution are known;
(b) only the initial distribution is known;
(c) both are unknown.

†*Address for correspondence*: Department of Statistics, University of Connecticut, Storrs, CT 06269-3120, USA.

 0035–9254/92/41389 $2.00

Details are given for binomial and normal models. Broemeling (1972) and Menzefricke (1981) also consider normal models. Bacon and Watts (1971), Ferreira (1975), Holbert and Broemeling (1977), Chin Choy and Broemeling (1980), Smith and Cook (1990) and Moen *et al.* (1985) look at changepoints in linear models. Booth and Smith (1982) and, in a slightly different fashion, West and Harrison (1986) consider time series models. Diaz (1982) and Hsu (1982) study sequences of gamma-type random variables. Raftery and Akman (1986) extend this to Poisson processes where the changepoint need not have occurred at an event time.

In our view wider use of the Bayesian framework has been impeded by the difficulties in computing required marginal posterior distributions of the changepoint and of the model parameters. In many of the above papers unsatisfying compromises have been made with regard to model specification and/or assumed knowledge of at least some of the model parameters to reduce the dimensionality of numerical integrations required to obtain marginal posterior distributions. The objective of this paper is to demonstrate that, for a very broad range of hierarchical changepoint models, such a compromise is not necessary. Rather, all desired marginal posterior densities can be obtained through a straightforward iterative Monte Carlo method. Needed random generation, whether or not conjugacy is assumed, can typically be carried out reasonably efficiently. We concede that for any particular situation there may be more efficient methods for obtaining desired marginal posteriors. By the same token, we can handle many situations which were previously inaccessible. Moreover the conceptual simplicity of our method may prove an attractive alternative to the analytical and/or numerical sophistication demanded by other methods. The subsequent development advances work reported in Gelfand and Smith (1990) and Gelfand *et al.* (1990).

In particular in Section 2 we formulate the hierarchical Bayes changepoint model. We clarify what distributions are sought and what distributions can be readily sampled. In Section 3 we briefly review the Gibbs sampler which underlies our iterative Monte Carlo method. We also show how the distribution of and expectations of arbitrary functions of the model parameters can be obtained. In Sections 4–6 we present a range of examples. As an elementary illustration in Section 4 we examine the British coal-mining accident data of Maguire *et al.* (1952) extended and corrected by Jarrett (1979). These data have been discussed in Worsley (1986), Raftery and Akman (1986) and Siegmund (1988). Our approach routinely handles the complete data set as well as a reduced version where 20% is treated as missing. In Section 5 we show that independent observations are not required to implement our methodology as we consider a change in transition matrix for a sequence of observations from a Markov chain. As a final example in Section 6 we study the changing regressions model, illustrated in the context of simple linear regression with a data set investigated by Bacon and Watts (1971).

2. Bayesian Formulation

The simplest formulation of the changepoint problem assumes f and g known densities with $Y_i \sim f(y)$, $i=1, \ldots, k$, $Y_i \sim g(Y)$, $i=k+1, \ldots, n$, with k unknown taking values in $\mathcal{K}_n = \{1, 2, \ldots, n\}$. Thus $k=n$ is interpreted as 'no change'. As a result with $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)$ the likelihood $L(\mathbf{Y}; k)$ becomes

$$L(\mathbf{Y}; k) = \prod_{i=1}^{k} f(Y_i) \prod_{i=k+1}^{n} g(Y_i) \tag{1}$$

from which, for instance, the maximum likelihood estimate (MLE) for k can be directly obtained. Distribution theory for the MLE is well discussed in the literature (see, for example, Hinkley (1970)). The Bayesian perspective is added by placing a prior density $\tau(k)$ on $\mathscr{K}_n$ whence the posterior density of $k \mid \mathbf{Y}$ becomes

$$L(\mathbf{Y}; k)\,\tau(k) \Big/ \sum_{k=1}^{n} L(\mathbf{Y}; k)\,\tau(k). \tag{2}$$

Features of density (2), e.g. mode, expectations, quantiles and credible intervals, are easy to obtain. The question of whether or not a change has occurred is addressed through the posterior odds for no change, $P(k=n \mid \mathbf{Y})/\{1-P(k=n \mid \mathbf{Y})\}$. It is easy to sample from density (2) since it is a discrete distribution on $\mathscr{K}_n$. Also the independence of the Y_i is not necessary provided that for each k we can write down the likelihood $L(\mathbf{Y}; k)$. Such a situation arises in the Markov chain example of Section 6. None-the-less, to simplify the notation and presentation, for the remainder of this section we assume that the Y_i are independent.

Suppose, as is more realistic, that the original and changed densities are unknown but that we assume a parametric family $f(Y \mid \theta)$ for the former and a (possibly different) parametric family $g(Y \mid \eta)$ for the latter. Then the likelihood $L(\mathbf{Y}; k, \theta, \eta)$ becomes

$$\prod_{i=1}^{k} f(Y_i \mid \theta) \prod_{i=k+1}^{n} g(Y_i \mid \eta).$$

Assuming a prior $\pi(\theta, \eta, k)$ on θ, η and k the joint distribution of data and parameters is

$$L(\mathbf{Y}; k, \theta, \eta)\,\pi(\theta, \eta, k). \tag{3}$$

Interest centres on the marginal posterior distribution of k as well as that of components of θ and of η.

The joint posterior distribution of θ, η and k given the data is proportional to expression (3). However, marginalization of expression (3) to obtain, for example, the distribution $k \mid \mathbf{Y}$ requires integration over θ and η which is generally a formidable analytic problem. Use of the Gibbs sampler, as discussed in the context of hierarchical Bayesian models in Gelfand and Smith (1990) and Gelfand *et al.* (1990), enables a straightforward sampling-based solution to such problems. Implementation requires sampling from the complete conditional distributions, $k \mid \mathbf{Y}, \theta, \eta$, $\theta \mid \mathbf{Y}, k, \eta$ and $\eta \mid \mathbf{Y}, k, \theta$. Each of these distributions is also proportional to expression (3). Moreover, given θ and η, $k \mid \mathbf{Y}, \theta, \eta$ is exactly of the form (2) and is, as noted earlier, straightforwardly sampled. Suppose that θ, η and k are assumed independent so that $\pi(\theta, \eta, k) = \lambda(\theta)\,\gamma(\eta)\,\tau(k)$. If λ is conjugate with f and γ is conjugate with g then $\theta \mid \mathbf{Y}, \eta, k$ does not depend on η and is merely the prior λ updated by the data $Y_1, \ldots, Y_k$ while $\eta \mid \mathbf{Y}, \theta, k$ is γ updated by the data $Y_{k+1}, \ldots, Y_n$. Since λ and γ are thus standard parametric families, sampling from these full conditionals is routine. Conjugate priors can be

made arbitrarily diffuse and have an attractive robustness property (see Morris (1983), p. 525).

If we abandon the assumptions of conjugacy for θ and for η and also the assumption of independence for θ, η and k we must sample from non-standardized densities of form (3). Random generation methods such as the ratio of uniforms method or perhaps the rejection method (see Devroye (1986) or Ripley (1987)) enable such sampling. In the present paper we confine ourselves to conjugate examples.

Extension to general hierarchical Bayes models is immediate. Suppose, for instance, that again θ and η are independent of each other and of k. Also suppose that the prior of θ takes the form $\lambda(\theta|\alpha)$ with conjugate prior $\rho(\alpha)$ on the hyperparameter α and the prior on η takes the form $\gamma(\eta|\beta)$ with conjugate prior $\phi(\beta)$ on the hyperparameter β. Now the joint distribution of the data and all parameters is

$$L(\mathbf{Y}; k, \theta, \eta)\,\tau(k)\,\lambda(\theta|\alpha)\,\rho(\alpha)\,\gamma(\eta|\beta)\,\phi(\beta). \tag{4}$$

Once again the full conditionals are easy to write down and to sample from. In particular:

(a) $k|\mathbf{Y}, \theta, \eta, \alpha, \beta$ does not depend on α or β and is exactly of form (2);
(b) $\theta|\mathbf{Y}, k, \eta, \alpha, \beta$ does not depend on η or β and is the prior λ updated by the data $Y_1, \ldots, Y_k$;
(c) $\eta|\mathbf{Y}, k, \theta, \alpha, \beta$ does not depend on θ or α and is the prior γ updated by the data $Y_{k+1}, \ldots, Y_n$;
(d) $\alpha|\mathbf{Y}, k, \theta, \eta, \beta$ does not depend on $\mathbf{Y}$, k, η or β and is the hyperprior ρ updated by θ;
(e) $\beta|\mathbf{Y}, k, \theta, \alpha, \eta$ does not depend on $\mathbf{Y}$, k, θ or α and is the hyperprior ϕ updated by η.

Again if the assumed conjugacy and parameter independence are relaxed the aforementioned random generation methods can still be used.

In all our examples we take a uniform prior on k. Since the event '$k=n$' denotes no change in some applications it might be sensible to place greater prior mass on this event. In some cases it may be of interest to obtain the posterior distribution of, say, θ given the data *and* that a change has occurred, i.e. $f(\theta|\mathbf{Y}, k\neq n)$. Using the identity

$$f((\theta|\mathbf{Y}) = f((\theta|\mathbf{Y}, k=n)\,P(k=n|\mathbf{Y}) + f((\theta|\mathbf{Y}, k\neq n)\,P(k\neq n|\mathbf{Y})$$

all distributions except the desired one can be estimated by using the Gibbs sampling approach whence we may solve for an estimate of $f((\theta|\mathbf{Y}, k\neq n)$.

3. Review of Gibbs Sampler

The Gibbs sampler is a Monte Carlo integration method which proceeds by a Markovian updating scheme. It was developed formally by Geman and Geman (1984) in the context of image restoration. In the statistical framework, Tanner and Wong (1987) used essentially this algorithm in their substitution sampling approach to missing data problems. More recently, Gelfand and Smith (1990) showed its applicability to general parametric Bayesian computations; the reader is referred to that paper for a more complete discussion of the method and its properties. To summarize the method briefly, suppose that we have a collection of p (possibly vector-valued) random variables $U_1, \ldots, U_p$ whose full conditional distributions, denoted

generically by $f(U_s | U_r,\ r \neq s)$, $s = 1, \ldots, p$, are available for sampling. Here, 'available' means that samples may be generated by some method, given values of the appropriate conditioning random variables. Under mild conditions (see Besag (1974)), these full conditional distributions uniquely determine the full joint distribution $f(U_1, \ldots, U_p)$, and hence all marginal distributions $f(U_s)$, $s = 1, \ldots, p$. The Gibbs sampler generates samples from the joint distribution as follows. Given an arbitrary starting set of values $U_1^{(0)}, \ldots, U_p^{(0)}$, we draw $U_1^{(1)}$ from $f(U_1 | U_2^{(0)}, \ldots, U_p^{(0)})$, then $U_2^{(1)}$ from $f(U_2 | U_1^{(1)}, U_3^{(0)}, \ldots, U_p^{(0)})$, and so on up to $U_p^{(1)}$ from $f(U_p | U_1^{(1)}, \ldots, U_{p-1}^{(1)})$ to complete one iteration of the scheme. After t such iterations we obtain $(U_1^{(t)}, \ldots, U_p^{(t)})$. Geman and Geman (1984) show under mild conditions that this p-tuple converges in distribution to a random observation from $f(U_1, \ldots, U_p)$ as $t \to \infty$. Hence replicating the entire process in parallel m times provides independent and identically distributed p-tuples $(U_{1j}^{(t)}, \ldots, U_{pj}^{(t)})$, $j = 1, \ldots, m$. These observations then can be used for estimation of any of the marginal densities that we desire. Gelfand and Smith (1990) recommend a density estimate of the form

$$\hat{f}(U_s) = \sum_{j=1}^{m} f(U_s | U_{rj}^{(t)}, r \neq s)/m. \tag{5}$$

Form (5) is a discrete mixture distribution and is in fact a Monte Carlo integration to accomplish the desired marginalization.

There may also be interest in a function of the parameters (see Section 4), i.e. a function of the U_i, say $W(U_1, \ldots, U_p)$. Each p-tuple $(U_{1j}^{(t)}, \ldots, U_{pj}^{(t)})$ provides an observed $W_j^{(t)} \equiv W(U_{1j}^{(t)}, \ldots, U_{pj}^{(t)})$ whose marginal distribution is approximately $f(W)$. A density estimate analogous to equation (5) can also be obtained. If U_s actually appears as an argument of W, the complete conditional density of $W | U_r$, $r \neq s$ can be obtained by univariate transformation from that of $U_s | U_r$, $r \neq s$.

In concluding this section we note that complete implementation of the Gibbs sampler requires that a determination of t be made and that, across iterations, choice(s) of m be specified. Some discussion of convergence monitoring is provided by Gelfand *et al.* (1990). In a challenging application experimentation with different settings for t and m will probably be necessary. We do not view this as a deterrent since random generation is generally inexpensive and since there may be no feasible alternative. In the subsequent examples t was at most 50 and we used $m = 100$ replications.

4. Poisson Process with Changepoint

As an illustration consider Bayesian analysis of a Poisson process with a changepoint. Such an analysis was recently given by Raftery and Akman (1986). Assuming conjugate priors they examine the collection of times between occurrences of the process allowing the time of change to be continuous. For chronologically ordered time intervals not necessarily of equal length we study the set of occurrence counts. We use a three-stage hierarchical model but assume that the changepoint occurs between intervals. Thus our model for the data is $Y_i \sim P_0(\theta t_i)$, $i = 1, \ldots, k$, $Y_i \sim P_0(\lambda t_i)$, $i = k+1, \ldots, n$. At the second stage we place independent priors over k, θ and λ: k discrete uniform $\mathcal{H}_n$, $\theta \sim G(a_1, b_1)$, and $\lambda \sim G(a_2, b_2)$ where G denotes the gamma distribution. At the third stage we take $b_1 \sim \mathrm{IG}(c_1, d_1)$ independent of

$b_2 \sim \text{IG}(c_2, d_2)$, where IG denotes the inverse gamma distribution, and assume that a_1, a_2, c_1, c_2, d_1 and d_2 are known. Our interest again lies in the marginal posterior distributions of k, θ and λ. The collection of complete conditional distributions is easily available as

$$\theta|\mathbf{Y}, \lambda, b_1, b_2, k \sim G\left\{a_1 + \sum_1^k Y_i, \left(\sum_1^k t_i + b_1^{-1}\right)^{-1}\right\},$$

$$\lambda|\mathbf{Y}, \theta, b_1, b_2, k \sim G\left\{a_2 + \sum_{k+1}^n Y_i, \left(\sum_{k+1}^n t_i + b_2^{-1}\right)^{-1}\right\},$$

$$b_1|\mathbf{Y}, \theta, \lambda, b_2, k \sim \text{IG}\{a_1 + c_1, (\theta + d_1^{-1})^{-1}\},$$

$$b_2|\mathbf{Y}, \theta, \lambda, b_1, k \sim \text{IG}\{a_2 + c_2, (\lambda + d_2^{-1})^{-1}\}$$

and

$$p(k|\mathbf{Y}, \theta, \lambda, b_1, b_2) = L(\mathbf{Y}; k, \theta, \lambda) \Big/ \sum_k L(\mathbf{Y}; k, \theta, \lambda)$$

where

$$L(\mathbf{Y}; k, \theta, \lambda) = \exp\left\{(\lambda - \theta)\sum_1^k t_i\right\}(\theta/\lambda)^{\Sigma_1^k Y_i}.$$

A variable of interest might be the ratio $R = \theta/\lambda$. Following the discussion near the end of Section 3 we transform from θ to R to obtain as full conditional distribution

$$R|\mathbf{Y}, \lambda, b_1, b_2, k \sim G\left\{a_1 + \sum_1^k Y_i, \lambda^{-1}\left(\sum_1^k t_i + b_1^{-1}\right)^{-1}\right\}.$$

A density estimate as in equation (1) can now be obtained.

The Gibbs sampler, in general, straightforwardly handles missing data. Such points can be treated as additional model parameters whose full conditional distributions are immediate. However, if the predictive distributions for such points are not of interest we would typically not perform the additional random generation required to incorporate them as model parameters. Rather, we would prefer to use the likelihood based solely on the observed data. In the present case, for example, we need only to set the associated Y and t for missing points to 0 in the likelihood and proceed as above.

A much analysed data set of intervals between British coal-mining disasters during the 112-year period 1851–1962 was gathered by Maguire *et al.* (1952), extended and corrected by Jarrett (1979). Frequentist changepoint investigations appear in Worsley (1986) and in Siegmund (1988) while Raftery and Akman (1986) apply their Bayesian model. We apply ours using yearly intervals converting Jarrett's table to the observed annual counts given in Table 1: $k \in \{1, 2, \ldots, 112\}$.

Raftery and Akman (1986) used a vague second-stage prior by taking $a_1 = a_2 = 0.5$ and $b_1 = b_2 = 0$. We make our model comparable by taking their values for a_1 and a_2 and choosing vague third-stage priors for b_1 and b_2, letting $c_1 = c_2 = 0$ and $d_1 = d_2 = 1$. Convergence of the algorithm was obtained after 15 iterations and again $m = 100$. Fig. 1 shows the density estimates for $k|\mathbf{Y}$ with the entire data set (full lines) and with

TABLE 1
British coal-mining disaster data by year, 1851–1962†

Year	*Count*	*Year*	*Count*	*Year*	*Count*	*Year*	*Count*
1851	4	1881	2	1911	0	1941	4
1852	5	1882	5	1912	1	1942	2
1853	4	1883	2	1913	1	1943	0
1854	1	1884	2	1914	1	1944	0
1855	0	1885	3	1915	0	1945	0
1856	4	1886	4	1916	1	1946	1
1857	3	1887	2	1917	0	1947	4
1858	4	1888	1	1918	1	1948	0
1859	0	1889	3	1919	0	1949	0
1860	6	1890	2	1920	0	1950	0
1861	3	1891	2	1921	0	1951	1
1862	3	1892	1	1922	2	1952	0
1863	4	1893	1	1923	1	1953	0
1864	0	1894	1	1924	0	1954	0
1865	2	1895	1	1925	0	1955	0
1866	6	1896	3	1926	0	1956	0
1867	3	1897	0	1927	1	1957	1
1868	3	1898	0	1928	1	1958	0
1869	5	1899	1	1929	0	1959	0
1870	4	1900	0	1930	2	1960	1
1871	5	1901	1	1931	3	1961	0
1872	3	1902	1	1932	3	1962	1
1873	1	1903	0	1933	1		
1874	4	1904	0	1934	1		
1875	4	1905	3	1935	2		
1876	1	1906	1	1936	1		
1877	5	1907	0	1937	1		
1878	5	1908	3	1938	1		
1879	3	1909	2	1939	1		
1880	4	1910	2	1940	2		

†Data from Maguire *et al.* (1952) corrected by Jarrett (1979).

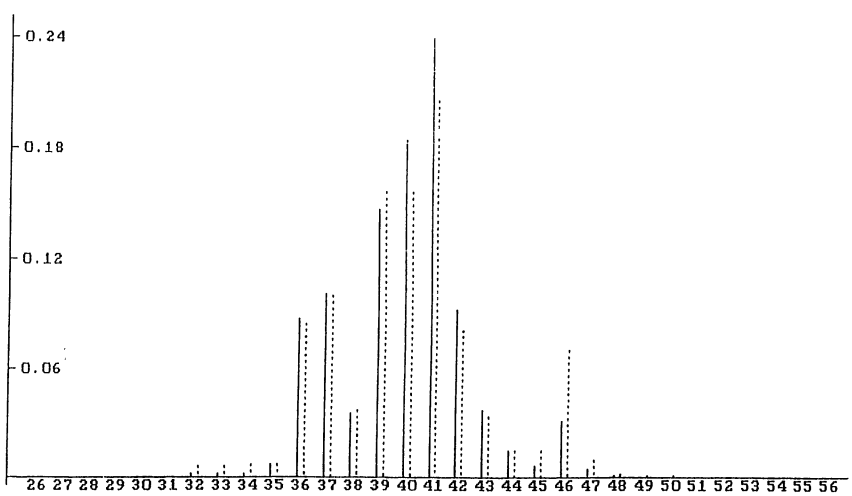

Fig. 1. Estimates of the density for $k \mid \mathbf{Y}$ for the coal-mine disaster data ($a_1 = a_2 = 0.5$; $c_1 = c_2 = 0$; $d_1 = d_2 = 1$): ———, full data; --------, 20% missing (reduced) data

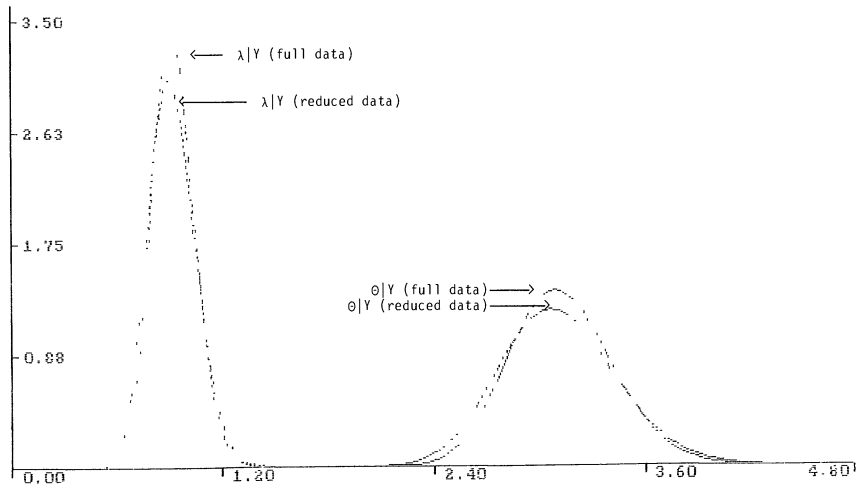

Fig. 2. Estimates of the densities for $\theta \mid \mathbf{Y}$ and $\lambda \mid \mathbf{Y}$ for the coal-mine disaster data ($a_1 = a_2 = 0.5$; $c_1 = c_2 = 0$; $d_1 = d_2 = 1$) with the full data and 20% missing (reduced) data

the data set with every fifth year deleted (broken lines). $k = 41$ is the posterior mode in both cases, and the three largest spikes are k equal to 39, 40 and 41, meaning that the change most probably occurred sometime between late 1889 and early 1892. Raftery and Akman obtained similar results—a posterior mode of March 10th, 1890, and a posterior median of August 27th, 1890. Also note that, in the missing data case, with no data at $k = 5j$ the density estimates are the same for $k = 5j$ and $k = 5j - 1, j = 1, \ldots,$ 22.

Fig. 2 displays the density estimates for $\theta \mid \mathbf{Y}$ and for $\lambda \mid \mathbf{Y}$ for the original and reduced data sets, and Fig. 3 does the same for $R \mid \mathbf{Y}$. For all three pairs of curves, as

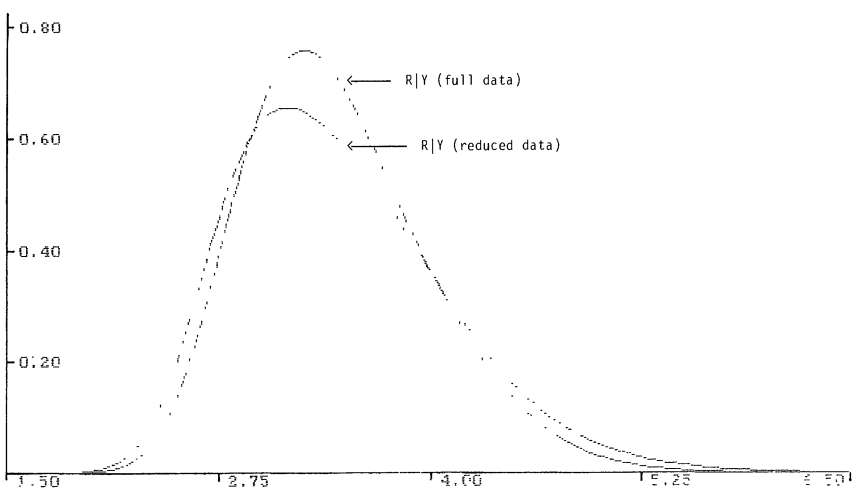

Fig. 3. Estimates of the density for $R \mid \mathbf{Y}$ for the coal-mine disaster data ($a_1 = a_2 = 0.5$; $c_1 = c_2 = 0$; $d_1 = d_2 = 1$) with the full data and 20% missing (reduced) data

expected the reduced data set has the greater spread. We see that systematically deleting 20% of the data has no effect on the posterior modes of θ and λ (3.06 and 0.89 respectively) and shrinks the posterior mode of R only slightly from 3.25 to 3.18. Raftery and Akman (1986) obtained 3.41 as the posterior mean of R with their model. As a final remark, we note that the posterior probability that $k = n$ is essentially 0, indicating very strong evidence for a change.

5. Markov Chain Changepoints

To our knowledge there is no previous literature examining the Markov chain changepoint problem from a Bayesian viewpoint. None-the-less the problem is extremely important, arising for instance in the analysis of spatial variations in base frequencies in deoxyribonucleic acid (Curnow and Kirkwood (1989), section 7) as well as in computer system user authentication contexts. Suppose then a sequence of a sequence of n observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from a process which is a p-state stationary Markov chain having either transition matrix A or precisely one change to a transition matrix B. The entries of A are $a_{ij} = P(Y_{t+1} = j \mid Y_t = i)$ whence $a_{ij} \geqslant 0$, $\Sigma_j a_{ij} = 1$; similarly for B with entries b_{ij}. We take independent Dirichlet priors on the rows of A yielding the prior

$$\lambda(A) = \prod_{i=1}^{p} D_{\lambda_i}(a_i)$$

where

$$D_{\lambda_i}(a_i) = \frac{\Gamma\left(\sum_{j=1}^{p} \lambda_{ij}\right)}{\prod_{j=1}^{p} \Gamma(\lambda_{ij})} \prod_{j=1}^{p} a_{ij}^{\lambda_{ij}}.$$

We do similarly for B taking

$$\gamma(B) = \prod_{i=1}^{p} D\gamma_i(b_i).$$

λ and γ offer a rich class of priors. In particular, special structure for A and B can be modelled through appropriate choices of the λ_i and γ_i. As before, we denote the prior on k by $\tau(k)$. The multinomial changepoint problem occurs as a special case when the Y_i are independent and $a_i = a$ and $b_i = b$.

We assume that A, B and k are independent whence the joint distribution of the data and parameters is

$$f(\mathbf{Y} \mid A, B, k)\, \tau(k)\, \lambda(A)\, \gamma(B) \tag{6}$$

where at $\mathbf{Y} = \mathbf{y} = (y_1, \ldots, y_n)$,

$$f(\mathbf{y}|A, B, k) = \begin{cases} \prod_{t=1}^{k-1} a_{y_t, y_{t+1}} \prod_{t=k}^{n-1} b_{y_t, y_{t+1}} \mu_0(y_1), & 2 \leqslant k \leqslant n-1, \\ \prod_{t=1}^{n-1} a_{y_t, y_{t+1}} \mu_0(y_1), & k = n, \\ \prod_{t=1}^{n-1} b_{y_t, y_{t+1}} \mu_0(y_1), & k = 1, \end{cases} \tag{7}$$

with μ_0 denoting the initial or starting state distribution. Using expressions (6) and (7) we see that the full posteriors are

$$A|\mathbf{Y}, B, k \sim \prod_{i=1}^{p} D_{\lambda_i + \mathbf{Z}_i}(a_i) \tag{8}$$

and

$$B|\mathbf{Y}, A, k \sim \prod_{i=1}^{p} D_{\gamma_i + \mathbf{Z}_i'}(a_i) \tag{9}$$

where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})$ with Z_{ij} the number of transitions from i to j within observations $Y_1, \ldots, Y_k$, $\mathbf{Z}_i' = (Z_{i1}', \ldots, Z_{ip}')$ with Z_{ij}' the number of transitions from i to j within observations $Y_k, \ldots, Y_n$ and

$$p(k|\mathbf{Y}, A, B) = f(\mathbf{Y}|A, B, k)\tau(k) \Big/ \sum_{k=1}^{n} f(\mathbf{Y}|A, B, k)\tau(k). \tag{10}$$

Since we are conditioning on all the data, and hence Y_1, μ_0 does not appear in posterior (8) or (9) and cancels in equation (10). Thus we need not worry about its specification.

A three-stage hierarchical model can be developed by placing priors on the λ_i and γ_i. Details are similar to those for the previous examples and are omitted. We note that work on hierarchical models for contingency tables as in, for example, Albert and Gupta (1982) is closely related.

We add the usual Bayesian caveat. The number of parameters involved in this analysis is $2p(p-1)+1$. Hence the smaller n is relative to $2p(p-1)+1$ the more the prior will drive the posterior densities. Other considerations may make it important to detect the changepoint as soon as possible rather than waiting to the end of the data sequence. Simple updating of the posterior given new data is not possible since with additional data the domain of k changes. If 'change' *versus* 'no change' is of primary concern, as an informal sequential procedure we might place a spike in the prior τ at $k=n$. For instance $\tau(n) = \frac{1}{2}$ implies prior indifference regarding a change in the first n observations. Use of a uniform prior would imply prior odds of $n-1$ to 1 for a change. Monitoring the posterior odds for no change, $P(k=n|\mathbf{Y})/\{1-P(K=n|\mathbf{Y})\}$, will reveal a downward trend to warn of change.

As an illustrative example we consider a three-state stationary Markov chain where

$$A = \begin{pmatrix} 0.7000 & 0.1500 & 0.1500 \\ 0.3333 & 0.3333 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 \end{pmatrix}, \qquad B = \begin{pmatrix} 0.3333 & 0.3333 & 0.3333 \\ 0.1500 & 0.7000 & 0.1500 \\ 0.3333 & 0.3333 & 0.3333 \end{pmatrix}.$$

Table 2 gives a sequence of 50 observations generated with a change after $k=35$.

$\tau(k)$ was taken to be uniform over $\{1, \ldots, 50\}$. The Dirichlet priors are taken to be

TABLE 2
Sequence of 50 observations for the three-state Markov chain example

1	1	2	2	1	1	1	1	1	1	2	3	3	2	3	2	1	1	2	1
1	3	3	1	1	1	3	2	1	1	1	1	1	3	2	2	3	1	2	2
2	2	2	2	2	3	2	3	2	2										

generalized uniforms prior, i.e. all λ_{ij}, $\gamma_{ij} = 1$. Turning to the marginal posteriors, in Fig. 4 we show for $k|\mathbf{Y}$ a comparison of the density at iteration 4 (full line) and at iteration 5 (broken line) using $m = 100$. We would not conclude convergence. In Fig. 5 we compare iterations 30 (full line) and 31 (broken line). Now a conclusion of convergence seems appropriate. In fact the density estimates have been roughly

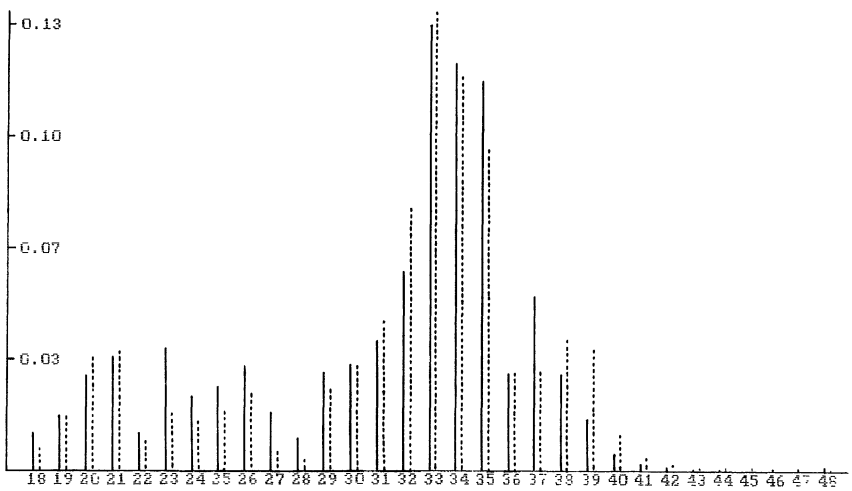

Fig. 4. Estimates of the density for $k|\mathbf{Y}$ for the Markov chain data, iterations 4 and 5

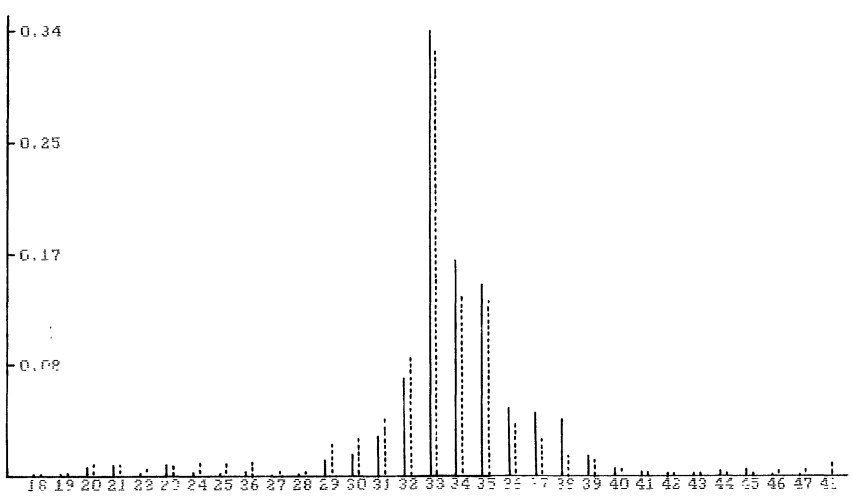

Fig. 5. Estimates of the density for $k|\mathbf{Y}$ for the Markov chain data, iterations 30 and 31

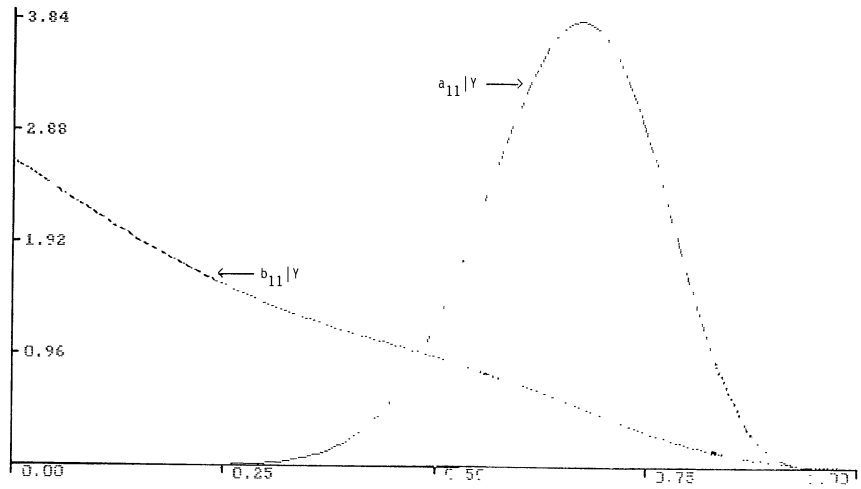

Fig. 6. Estimates of the densities for $a_{11}|\mathbf{Y}$ and $b_{11}|\mathbf{Y}$ for the Markov chain data

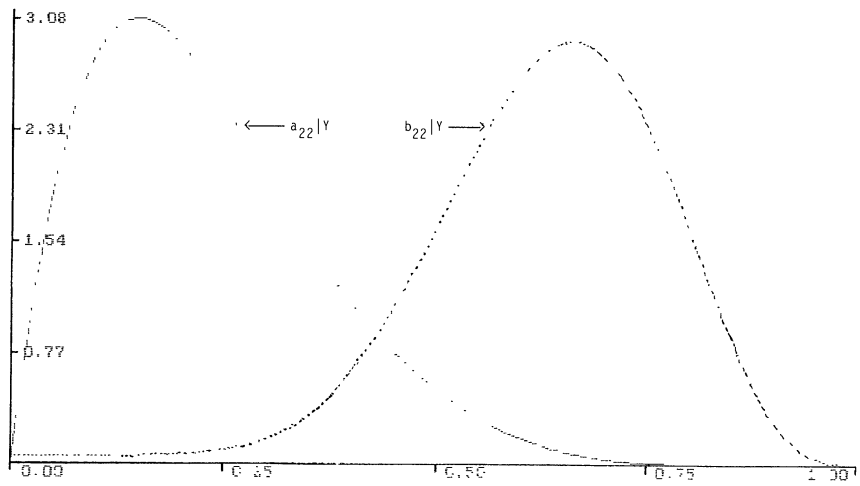

Fig. 7. Estimates of the densities for $a_{22}|\mathbf{Y}$ and $b_{22}|\mathbf{Y}$ for the Markov chain data

this stable since iteration 20. The marginal posterior for k has mode at 33 with roughly 60% of the mass on the set $\{33, 34, 35\}$. In Fig. 6 we show the marginal posteriors for $a_{11}|\mathbf{Y}$ and $b_{11}|\mathbf{Y}$. In Fig. 7 we show the marginal posteriors for $a_{22}|\mathbf{Y}$ and $b_{22}|\mathbf{Y}$. For both a_{11} and b_{22}, modes and concentration of mass agree with the 'true' a_{11} and b_{22} respectively. Interestingly, for b_{11} the marginal posterior is one tailed. This arises because for the observed sequence there are no transitions from state 1 to state 1 after $k=33$.

6. Changing Linear Regression Models

Bayesian analysis of changing linear models includes the work of Bacon and Watts (1971), Ferreira (1975), Holbert and Broemeling (1977), Chin Choy and Broemeling

(1980), Smith and Cook (1980) and Moen *et al.* (1985). Typically this work involves simplified models and/or considerable analytical effort. By contrast the Gibbs sampler enables analysis of complex models while demanding little mathematical and computational expertise from the user.

Consider a three-stage hierarchical simple linear regression model, where at the first stage $Y_i \sim N(\alpha_1+\beta_1 x_i, \sigma_1^2)$, $i=1, \ldots, k$, $Y_i \sim N(\alpha_2+\beta_2 x_i, \sigma_2^2)$, $i=k+1, \ldots, n$. At the second stage we take $\theta_1 = (\alpha_1, \beta_1)^{\mathrm{T}}$, $\theta_2 = (\alpha_2, \beta_2)^{\mathrm{T}}$ independent $N(\theta_0, \Sigma)$ and σ_1^2, σ_2^2 independent $\mathrm{IG}(a_0, b_0)$. Again we assume that k follows a discrete uniform distribution on $\mathscr{K}_n$.

At the third stage, we take the independent normal–Wishart form $\theta_0 \sim N(\mu, C)$, $\Sigma^{-1} \sim W\{(\rho V)^{-1}, \rho\}$. Thus we have a 13-parameter problem with known constants μ, C, V, ρ, a_0 and b_0.

For a given k define $B_i^{(k)} = (\sigma_i^{-2}\mathbf{X}_i^{(k)\mathrm{T}}\mathbf{X}_i^{(k)} + \Sigma^{-1})^{-1}$, $b_i^{(k)} = \sigma_i^{-2}\mathbf{X}_i^{(k)\mathrm{T}}\mathbf{Y}_i^{(k)} + \Sigma^{-1}\theta_0$, $i=1, 2$, where

$$\mathbf{X}_1^{(k)} = \begin{pmatrix} 1 & \ldots & 1 \\ X_1 & \ldots & X_k \end{pmatrix}, \qquad \mathbf{X}_2^{(k)} = \begin{pmatrix} 1 & \ldots & 1 \\ X_{k+1} & \ldots & X_n \end{pmatrix}^{\mathrm{T}},$$

$\mathbf{Y}_1^{(k)} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{Y}_2^{(k)} = (Y_{k+1}, \ldots, Y_n)^{\mathrm{T}}$. Let $\Delta = (2\Sigma^{-1} + C^{-1})^{-1}$. Then using standard distribution theory we obtain the following full conditional distributions: $\theta_i \sim N(B_i^{(k)} b_i^{(k)}, B_i^{(k)})$, $i=1, 2$;

$$\sigma_1^2 \sim \mathrm{IG}\left(a_0+\frac{k}{2}, \left\{\frac{1}{2}(\mathbf{Y}_1^{(k)}-\mathbf{X}_1^{(k)}\theta_1)^{\mathrm{T}}(\mathbf{Y}_1^{(k)}-\mathbf{X}_1^{(k)}\theta_1)+\frac{1}{b_0}\right\}^{-1}\right);$$

$$\sigma_2^2 \sim \mathrm{IG}\left(a_0+\frac{n-k}{2}, \left\{\frac{1}{2}(\mathbf{Y}_2^{(k)}-\mathbf{X}_2^{(k)}\theta_2)^{\mathrm{T}}(\mathbf{Y}_2^{(k)}-\mathbf{X}_2^{(k)}\theta_2)+\frac{1}{b_0}\right\}^{-1}\right);$$

$$\theta_0 \sim N(\Delta\{\Sigma^{-1}(\theta_1+\theta_2)+C^{-1}\mu\}, \Delta);$$

$$\Sigma^{-1} \sim W\left(\left\{\sum_1^2 (\theta_i-\theta_0)(\theta_i-\theta_0)^{\mathrm{T}}+\rho V\right\}^{-1}, \rho+2\right)$$

and

$$k \sim \rho(k \mid \mathbf{Y}, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \theta_0, \Sigma) = L(\mathbf{Y}; k, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \Big/ \sum_{\mathscr{K}_n} L(\mathbf{Y}; k, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2)$$

where

$$L(\mathbf{Y}; k, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) = \exp\left\{-\sum_{i=1}^{2} \frac{1}{2\sigma_i^2}(\mathbf{Y}_i^{(k)}-\mathbf{X}_i^{(k)}\theta_i)^{\mathrm{T}}(\mathbf{Y}_i^{(k)}-\mathbf{X}_i^{(k)}\theta_i)\right\} \Big/ \sigma_1^k \sigma_2^{n-k}.$$

Simulation from the Wishart distribution is easily done by using the algorithm of Odell and Feiveson (1966) as outlined for the 2×2 case in Gelfand *et al.* (1990).

We apply the Gibbs sampler thus defined to the data in Table 3, which come from Bacon and Watts (1971), p. 530. In these data, X represents the logarithm of the flow rate of water down an inclined channel (grams per centimetre per second) and Y represents the logarithm of the height of the stagnant surface layer (centimetres) for different surfactants.

The data seem to indicate a decreasing linear trend that becomes more steeply

TABLE 3
Stagnant band height data†

X	Y	X	Y	X	Y	X	Y
0.11	0.44	0.11	0.43	0.34	0.25	−1.08	0.99
−0.80	0.90	0.11	0.43	1.19	−0.65	−1.08	1.03
0.01	0.51	−0.63	0.81	0.59	−0.01	0.44	0.13
−0.25	0.65	−0.63	0.83	0.85	−0.30	0.34	0.24
−0.25	0.67	−1.39	1.12	0.85	−0.33	0.25	0.30
−0.12	0.60	−1.39	1.12	0.99	−0.46		
−0.12	0.59	0.70	−0.13	0.99	−0.43		
−0.94	0.92	0.70	−0.14	0.25	0.33		

†X = log(flow rate), Y = log(band height), from Bacon and Watts (1971).

decreasing for $X > 0$. We apply our model using a vague prior for the σ_i^2, $a_0 = 0.1$ and $b_0 = 100$, along with a vague third-stage prior for θ_0, all entries of μ and C^{-1} equal to 0. For Σ^{-1} we compare two different Wishart priors, the first somewhat informative

$$\left(\rho = 4,\ V = \begin{pmatrix} 0.001 & 0 \\ 0 & 0.3 \end{pmatrix}\right),$$

and the second more vague

$$\left(\rho = 2,\ V = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right).$$

Convergence was obtained within 50 iterations and densities are drawn at the 51st with $m = 100$. Fig. 8 plots the estimated marginal posterior for k with the informative (full line) and vague (broken line) priors on Σ^{-1}. Fig. 9 does the same for β_1 and β_2, the slopes before and after the change (the β_2 estimates are on the left-hand side). Clearly with non-informative priors on θ_0, σ_1^2 and σ_2^2, changing the prior on Σ^{-1} can have a rather marked effect on the results. For example, k's posterior distribution is much more diffuse with the vague Wishart prior. Our intent here is not to claim that one or the other solution is correct, but rather to emphasize how simple it is using the Gibbs sampler for the data analyst to investigate prior robustness interactively.

The estimated posterior modes of β_1 for the informative and vague priors are -0.42 and -0.44 respectively, and for β_2 they are -1.01 and -0.98 respectively. Though Bacon and Watts (1971) used a different parameterization, a somewhat less elaborate model, vague prior specification and treated k as continuous, their results are quite comparable with ours: their joint posterior obtains its global maximum at $\beta_1 = -0.44$, $\beta_2 = -1.02$, with the change estimated to have occurred at $X = 0.054$. This corresponds to $k = 13$, which is the posterior median that we obtain under either of our prior specifications.

As a final remark, in certain applications we know that, if there has been a change, the change must be in a certain direction. We would want to incorporate this prior knowledge into our analysis. This entails placing order constraint(s) on the parameters, which makes required integrations much more difficult, perhaps impossible. The Gibbs sampler can again be used to overcome this difficulty; see

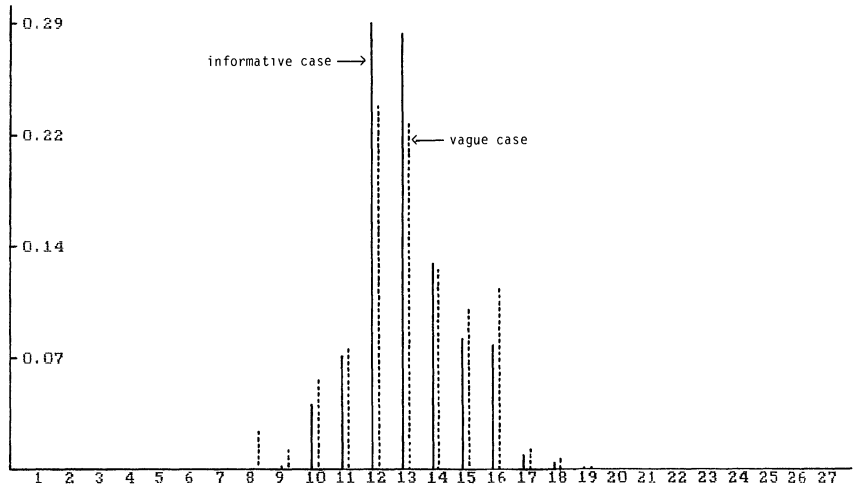

Fig. 8. Estimates of the density for $k|\mathbf{Y}$ for the Bacon and Watts (1971) data with $m = 100$, $a_0 = 0.1$, $b_0 = 100$, informative case $\rho = 4$,

$$V = \begin{pmatrix} 0.001 & 0 \\ 0 & 0.3 \end{pmatrix},$$

and vague case $\rho = 2$,

$$V = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

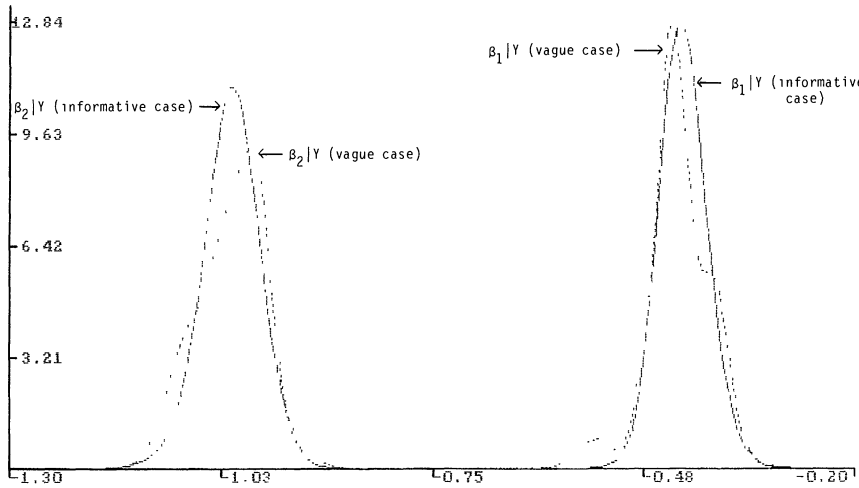

Fig. 9. Estimates of the densities for $\beta_1|\mathbf{Y}$ and $\beta_2|\mathbf{Y}$ for the Bacon and Watts (1971) data with $m = 100$, $a_0 = 0.1$, $b_0 = 100$, informative case $\rho = 4$,

$$V = \begin{pmatrix} 0.001 & 0 \\ 0 & 0.3 \end{pmatrix},$$

and vague case $\rho = 2$,

$$V = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Gelfand *et al.* (1991) for a general discussion of order-restricted parameter problems.

References

Albert, J. H. and Gupta, A. K. (1982) Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Statist.*, **10**, 1261–1268.

Bacon, D. W. and Watts, D. G. (1971) Estimating the transition between two intersecting straight lines. *Biometrika*, **58**, 525–534.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–226.

Booth, N. B. and Smith, A. F. M. (1982) A Bayesian approach to retrospective identification of change points. *J. Econ.*, **19**, 7–22.

Broemeling, L. (1972) Bayesian procedures for detecting a change in a sequence of random variables. *Metron*, **30**, 1–14.

Chernoff, H. and Zacks, S. (1964) Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.*, **35**, 999–1018.

Chin Choy, J. and Broemeling, L. (1980) Some Bayesian inferences for a changing linear model. *Technometrics*, **22**, 71–78.

Curnow, R. N. and Kirkwood, T. B. L. (1989) Statistical analysis of deoxyribonucleic acid sequence data—a review. *J. R. Statist. Soc.* A, **152**, 199–220.

Devroye, L. (1986) *Non-uniform Random Variate Generation*. New York: Springer.

Diaz, J. (1982) Bayesian detection of a change of scale parameter in sequences of independent gamma random variables. *J. Econ.*, **19**, 23–29.

Ferreira, P. E. (1975) A Bayesian analysis of a switching regression model: a known number of regimes. *J. Am. Statist. Ass.*, **70**, 370–374.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Am. Statist. Ass.*, **85**, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gelfand, A. E., Smith, A. F. M. and Lee, T. M. (1991) Bayesian analysis of constrained parameter and truncated data problems. *J. Am. Statist. Ass.*, to be published.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.

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

Hinkley, D., Chapman, P. and Runger, G. (1980) Change point problems. *Technical Report 382*. University of Minnesota, Minneapolis.

Holbert, D. and Broemeling, L. D. (1977) Bayesian inference related to shifting sequences and two-phase regression. *Communs Statist.* A, **6**, 265–275.

Hsu, D. A. (1982) A Bayesian robust detection of shift in the risk structure of stock market returns. *J. Am. Statist. Ass.*, **77**, 29–39.

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

Maguire, B. A., Pearson, E. S. and Wynn, A. H. A. (1952) The time intervals between industrial accidents. *Biometrika*, **38**, 168–180.

Menzefricke, U. (1981) A Bayesian analysis of a change in the precision of a sequence of independent normal random variables at an unknown time point. *Appl. Statist.*, **30**, 141–146.

Moen, D. H., Salazar, D. and Broemeling, L. D. (1985) Structural changes in multivariate regression models. *Communs Statist.* A, **14**, 1757–1768.

Morris, C. N. (1983) Natural exponential families with quadratic variance functions: statistical theory. *Ann. Statist.*, **11**, 515–529.

Odell, P. L. and Feiveson, A. H. (1966) A numerical procedure to generate a sample covariance matrix. *J. Am. Statist. Ass.*, **61**, 198–203.

Raftery, A. E. and Akman, V. E. (1986) Bayesian analysis of a Poisson process with a change-point. *Biometrika*, **73**, 85–89.

Ripley, B. (1987) *Stochastic Simulation*. New York: Wiley.

Shiryayev, A. N. (1963) On optimum methods in quickest detection problems. *Theory Probab. Applic.*, **8**, 22–46.

Siegmund, D. (1986) Boundary crossing probabilities and statistical applications. *Ann. Statist.*, **14**, 361–404.

———(1988) Confidence sets in change point problems. *Int. Statist. Rev.*, **56**, 31–48.

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

Smith, A. F. M. and Cook, D. G. (1980) Straight lines with a change-point: a Bayesian analysis of some renal transplant data. *Appl. Statist.*, **29**, 180–189.

Tanner, M. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528–550.

West, M. and Harrison, P. J. (1986) Monitoring and adaptation in Bayesian forecasting models. *J. Am. Statist. Ass.*, **81**, 741–750.

Wolfe, D. A. and Schechtman, E. (1984) Nonparametric statistical procedures for the change point problem. *J. Statist. Planng Inf.*, **9**, 389–396.

Worsley, K. J. (1986) Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, **73**, 91–104.

Zacks, S. (1983) Survey of classical and Bayesian approaches to the change point problem: fixed sample and sequential procedures of testing and estimation. In *Recent Advances in Statistics: Herman Chernoff Festschrift*, pp. 245–269. New York: Academic Press.