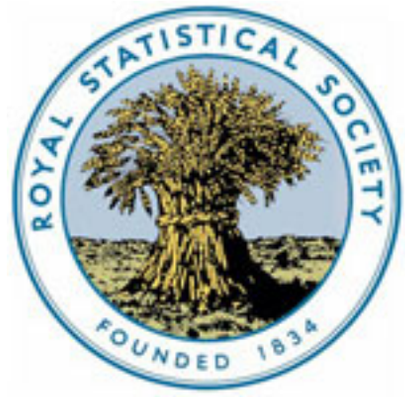

Appl. Statist. (1994)
43, *No.* 1, *pp.* 159–178

Bayesian Retrospective Multiple-changepoint Identification

By D. A. STEPHENS†

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

[Received September 1991. Final revision September 1992]

SUMMARY

Changepoint identification is important in many data analysis problems, such as industrial control and medical diagnosis—given a data sequence, 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 or parameters driving the system. For long data sequences, however, analysis (especially in the multiple-changepoint case) can become computationally prohibitive, and for complex non-linear models analytical and conventional numerical techniques are infeasible. We discuss the use of a sampling-based technique, the Gibbs sampler, in multiple-changepoint problems and demonstrate how it can be used to reduce the computational load involved considerably. Also, often it is reasonable to presume that the data model itself is continuous with respect to time, i.e. continuous at the changepoints. This necessitates a continuous parameter representation of the changepoint problem, which also leads to computational difficulties. We demonstrate how inferences can be made readily in such problems by using the Gibbs sampler. We study three examples: a simple discrete two-changepoint problem based on a binomial data model; a continuous switching linear regression problem; a continuous, non-linear, multiple-changepoint problem.

Keywords: Discrete and continuous changepoint models; Gibbs sampler; Multiple changepoint models

1. Introduction

Changepoint-type problems arise in many practical instances, e.g. signal processing, industrial system control, economics, medicine etc. Typically, chronologically ordered data are collected over a period of time during which there is known (or suspected) to have been a change in the underlying data generation process. Interest then lies in, retrospectively, making inference about the time or position in the sequence that this change occurred. This simple problem can be generalized to incorporate notions of multiple changes in the system.

For illustration of the practical use of changepoint models, consider the following real data examples that will be studied at greater length in the later sections of this paper. Fig. 1 depicts a data set studied by Smith (1980), known as the Lindisfarne scribes data. The data, denoted (Y_{1i}, Y_{2i}), $i=1, \ldots, 13$, in Table 1, are the number of occurrences of two types of pronoun ending observed in 13 chronologically ordered mediaeval manuscripts. The set of 13 documents is believed to be the work of more than one author, as the proportion of each ending in individual

†*Address for correspondence*: Department of Mathematics, Imperial College of Science, Technology and Medicine, Huxley Building, 180 Queen's Gate, London, SW7 2BZ, UK.

 0035–9254/94/43159

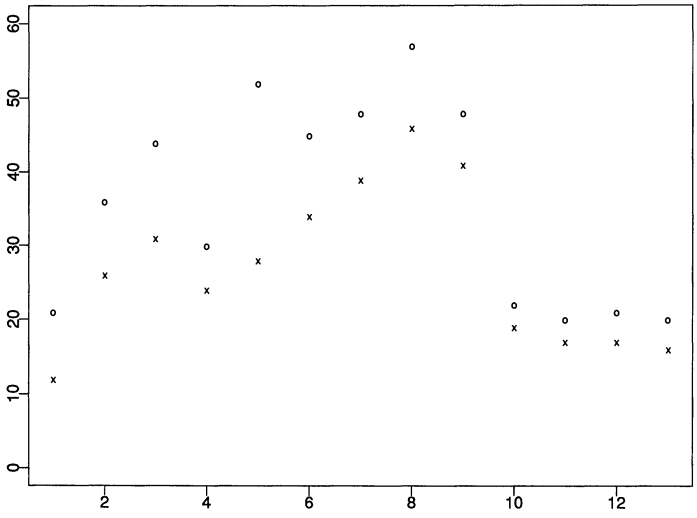

Fig. 1. Scribes data: ×, number ending one; ○, total number

TABLE 1
Example 1—scribes data

	No. of occurrences for the following values of i:												
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
Y_{i1}	12	26	31	24	28	34	39	46	41	19	17	17	16
Y_{i2}	9	10	13	6	24	11	9	11	7	3	3	4	4
Total N_i	21	36	44	30	52	45	48	57	48	22	20	21	20

documents appears to vary over the sequence. Therefore, assuming that the documents can be categorized into temporally contiguous phases, with each phase having a distinctive underlying proportion of ending one, say, and corresponding to a different scribe, it is clear that a discrete, possibly multiple-changepoint model is appropriate for these data. A suitable sampling distribution would be a binomial model, with each phase characterized by a different 'success' probability. Fig. 2 depicts two data sets arising in a medical example studied by Smith and Cook (1980). Patients A and B have undergone renal transplants, but the transplanted kidneys have subsequently been rejected. The response variable in this example, Y_t, is a (reciprocal, body weight adjusted) measurement of a substance present in the body, serum-creatinine, the amount of which corresponds directly to healthy renal function, at time t, recorded at daily intervals until up to 10 days after surgery. On the basis of these data, it is of clinical interest to assess retrospectively when rejection occurred (we shall not study here the prospective version of this problem in which it is required to detect the rejection 'on line'). Although rejection is probably not instantaneous, it occurs over a short time, either side of which the renal function can be presumed constant, and so a two-phase changepoint model

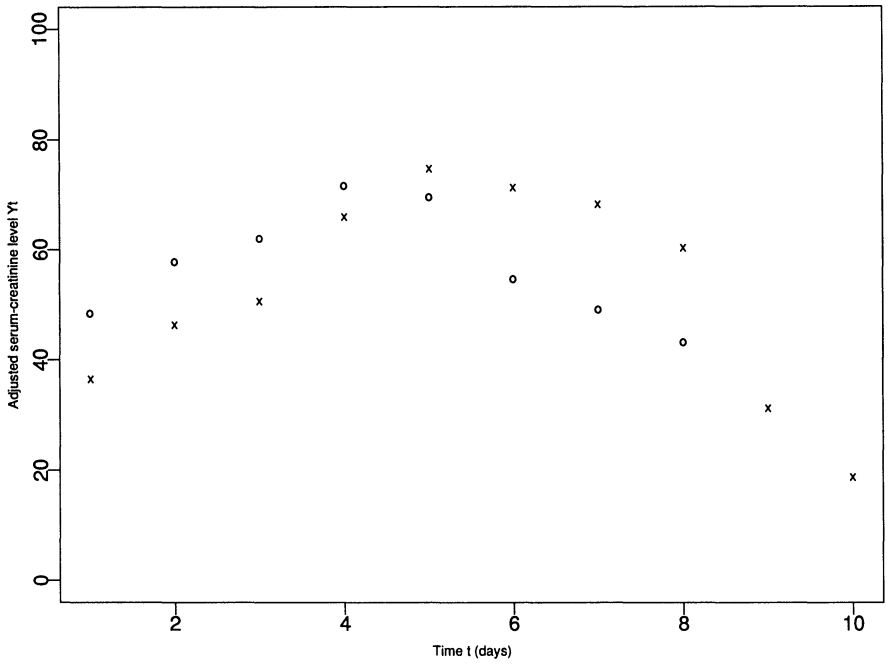

Fig. 2. Renal transplant data: ○, patient A; ×, patient B

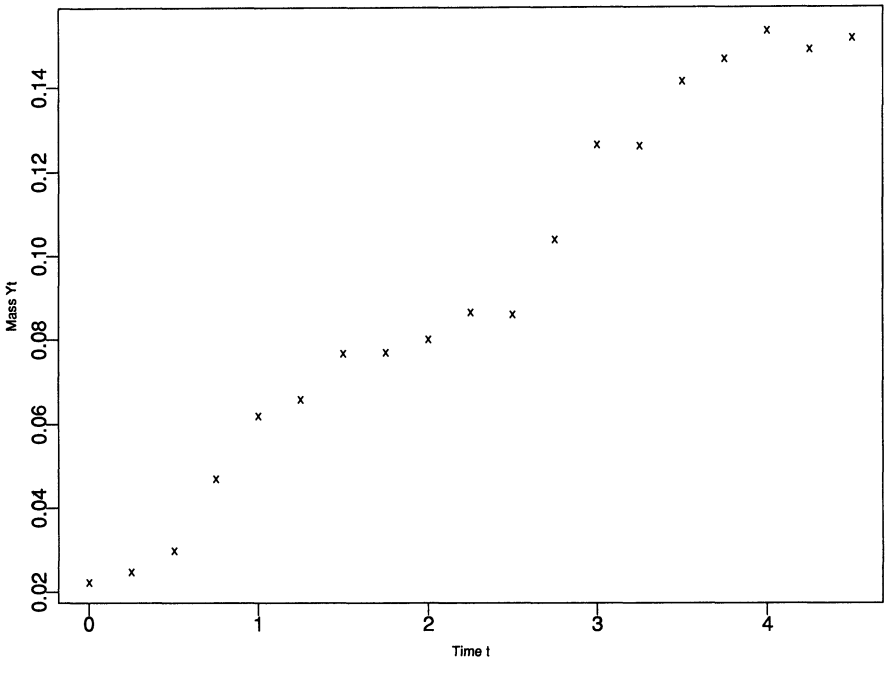

Fig. 3. Growth data

is adequate. Smith and Cook (1980) use a switching straight line model and recognize that, as there is no possibility of a sudden jump or drop in the level of serum-creatinine, the response curve should be modelled as being continuous at the changepoint. We shall see later how this continuity assumption necessarily leads to a continuous parameter representation of the changepoint problem, rather than the more common discrete representation. Finally, Fig. 3 depicts data from Schulze (1984) from a biological experiment to record the growth in terms of mass of bacteria in a nutritive medium. It is well known that there are four growth phases; in the first and third, mass increases exponentially in time, whereas, in the second and fourth, mass is constant. As in our second example, the nature of the problem indicates that the growth curve must be continuous. Here, therefore, a three-changepoint, continuous parameterization model is required.

In this paper, we briefly review previous work in this area and then study a discrete multiple-changepoint problem, a continuous single-changepoint problem and a continuous multiple-changepoint problem. We give an illustration of the exact Bayesian solution to multiple-changepoint problems via the scribes example but then demonstrate how conventional Bayesian analysis can become infeasible owing to computational complexity. We then describe the implementation of a computational technique, the Gibbs sampler algorithm, that can be used to facilitate Bayesian analysis in awkward problems. Finally, we analyse the data sets described above.

2. Single- and Multiple-changepoint Analysis

Several approaches to the changepoint problem have been published, including those based on nonparametric (Pettitt, 1980; Hinkley, 1971) or likelihood (Hinkley, 1970) formalisms. A Bayesian approach has been adopted by, for example, Chernoff and Zacks (1964), Broemeling (1974), Smith (1975) and Booth and Smith (1982). A Gibbs sampling approach to Bayesian inference for single-changepoint problems was presented by Carlin *et al.* (1992). In this paper, we extend the analysis of Carlin *et al.* (1992) by considering multiple-changepoint models, continuity constraints and non-linear models. We now introduce the relevant notation and terminology.

2.1. *Bayesian Retrospective Changepoint Identification*

We begin with a conventional discrete parameter representation and adopt the following notation. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a realization of the sequence of random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ of length n. Let $\boldsymbol{\theta}$ be the vector of parameters of the sampling distribution and $\boldsymbol{\psi}$ be a vector of hyperparameters appearing in the specification of the prior distribution for $\boldsymbol{\theta}$, denoted $[\boldsymbol{\theta}|\boldsymbol{\psi}]$ (following the square-bracket notation of Gelfand and Smith (1990)). The random variables $Y_1, \ldots, Y_n$ have a *changepoint* at r ($1 \leqslant r \leqslant n$) if

$$Y_1, \ldots, Y_r \sim [Y_i|\theta_1]_1,$$
$$Y_{r+1}, \ldots, Y_n \sim [Y_i|\theta_2]_2$$

where the sampling densities for the Y_i satisfy

$$[\ |\theta_1]_1 \neq [\ |\theta_2]_2.$$

Here, our emphasis will be on *retrospective* changepoint identification, i.e., given $\mathbf{Y}=\mathbf{y}$, our objective is, primarily, to make inferences about the unknown changepoint position r. Inference will be made via the posterior distribution of r, denoted by $[r|\mathbf{Y}, \psi]$. From Bayes's theorem, we have that

$$[r|\mathbf{Y}, \psi] \propto [\mathbf{Y}|r, \psi][r]. \tag{1}$$

The first term on the right-hand side of expression (1) is the likelihood function for $\mathbf{Y}$ integrated over the prior for θ, namely

$$[\mathbf{Y}|r, \psi] = \int [\mathbf{Y}|r, \theta, \psi][\theta|r, \psi]$$

if θ is wholly or partially unknown, and simply as the likelihood itself if θ is completely known (in which case we identify ψ as θ).

By assumption, the conditional distribution of $\mathbf{Y}$ given r, θ and ψ is independent of ψ. We shall also assume that $Y_1, \ldots, Y_n$ are conditionally independent given θ and regard θ as independent of r *a priori*. Thus, from expression (1), the posterior distribution of r is given by

$$[r|\mathbf{Y}, \psi] \propto \int \prod_{i=1}^{n} [Y_i|r, \theta][\theta|\psi][r]. \tag{2}$$

Finally, we often report some estimate of r, $\hat{r}$, say, obtained via $[r|\mathbf{Y}, \psi]$, rather than merely $[r|\mathbf{Y}, \psi]$ itself, such as the posterior mode. In the single-changepoint case, $[r|\mathbf{Y}, \psi]$ is simply a univariate, n-valued discrete distribution and thus will be easily calculable, with straightforward optimization, moment calculation, etc.

The Bayesian approach to the equivalent problems associated with multiple-changepoint sequences is identical with that above, i.e. we would make inference via $[r_1, r_2, \ldots, r_k|\mathbf{Y}, \psi]$ where

$$[r_1, r_2, \ldots, r_k|\mathbf{Y}, \psi] \propto \int \prod_{i=1}^{n} [Y_i|r_1, r_2, \ldots, r_k, \theta][\theta|\psi][r_1, r_2, \ldots, r_k], \tag{3}$$

where, without loss of generality, we assume that $r_1 < r_2 < \ldots < r_k$.

Finally, let M_k denote the model under which we assume precisely k changepoints. In most problems, we shall be uncertain about the number of changepoints (if any) that have occurred in any given sequence, but we shall be able to specify a prior distribution over a finite set of M_ks. In this case, the marginal posterior probabilities of a collection of k changepoints are given by expression (3), but with a multiplicative factor, the prior probability of M_k, included, and with the normalization constant being the summation over all possible k-changepoint models for each M_k. There is an implicit conditioning on M_1 and M_k in expressions (2) and (3) respectively, which we have previously suppressed notationally. Usually, the number of such models considered is small. Where more than one M_k is to be considered, it is generally very difficult to carry out a formal choice between the models, except purely in terms of posterior probabilities, as explained in the next section.

In this paper we shall only be concerned with problems where the changepoint(s) arise as the result of parametric rather than distributional changes in the data generation process, i.e. where $[\ |\ ]_1 = [\ |\ ]_2$, but $\theta_1 \neq \theta_2$. In any practical

problem, interest will lie in proposing various likelihood-prior combinations and examining the resulting posterior forms.

2.2. *Example 1—Lindisfarne Scribes Data*

Smith (1980) reviewed the analysis of the set of data given in Table 1 and described in Section 1. As indicated, these data were assumed to have arisen (independently) from a binomial multiple-changepoint sequence. Smith found evidence to suggest that at least two changepoints were present, and so we restrict attention to model M_2. The computation of posterior probabilities for other such models proceeds analogously, thus permitting model choice by means of posterior probability (provided that the priors used are proper)—we merely weight the probability for each k-tuple by $p(M_k)$ for the various k of interest and report the k and k-tuple having highest (joint) posterior probability.

Conditional on model M_2, in which we presume precisely two changepoints, this problem in our notation is formulated as follows: we denote the unknown changepoint positions r_1 and r_2, recall that $r_1 < r_2$ and replace $\boldsymbol{\theta}$ by $(\theta_1, \theta_2, \theta_3)$, so that

$$Y_{i1} | \theta_1, \theta_2, \theta_3, N_i \sim \begin{cases} \text{bin}(\theta_1, N_i) & i = 1, \ldots, r_1, \\ \text{bin}(\theta_2, N_i) & i = r_1 + 1, \ldots, r_2, \\ \text{bin}(\theta_3, N_i) & i = r_2 + 1, \ldots, 13. \end{cases}$$

Smith uses independent beta(0, 0) priors for the unknown $(\theta_1, \theta_2, \theta_3)$, i.e.

$$[\theta_1, \theta_2, \theta_3 | \psi] \propto \frac{1}{\theta_1\theta_2\theta_3(1-\theta_1)(1-\theta_2)(1-\theta_3)}, \tag{4}$$

and thus, via the two-changepoint version of expression (3), assuming a uniform prior on all possible pairs of values for the two changepoints, the joint posterior distribution for r_1 and r_2 can be seen to be given by

$$[r_1, r_2 | \mathbf{Y}, \mathbf{N}] \propto \frac{\Gamma(S_{11}+1)\Gamma(S_{12}+1)}{\Gamma(S_{11}+S_{12}+2)} \frac{\Gamma(S_{21}+1)\Gamma(S_{22}+1)}{\Gamma(S_{21}+S_{22}+2)} \frac{\Gamma(S_{31}+1)\Gamma(S_{32}+1)}{\Gamma(S_{31}+S_{32}+2)}$$
$$1 \leqslant r_1 < r_2 < 13 \tag{5}$$

TABLE 2
Example 1—exact joint posterior distribution under the two-changepoint model

r_2	*Probabilities for the following values of r_1:*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
2	0.001	–	–	–	–	–	–	–	–	–	–
3	0.001	0.000	–	–	–	–	–	–	–	–	–
4	0.000	0.000	0.000	–	–	–	–	–	–	–	–
5	0.065	0.029	0.035	0.328	–	–	–	–	–	–	–
6	0.061	0.023	0.019	0.036	0.048	–	–	–	–	–	–
7	0.014	0.005	0.003	0.003	0.030	0.020	–	–	–	–	–
8	0.006	0.002	0.001	0.001	0.029	0.018	0.004	–	–	–	–
9	0.001	0.000	0.000	0.000	0.022	0.016	0.003	0.001	–	–	–
10	0.001	0.000	0.000	0.000	0.022	0.018	0.003	0.001	0.000	–	–
11	0.000	0.000	0.000	0.000	0.026	0.022	0.004	0.002	0.000	0.000	–
12	0.001	0.000	0.000	0.000	0.036	0.029	0.005	0.002	0.000	0.000	0.000

where $S_{1j} = \Sigma_{i=1}^{r_1} Y_{ij}$, $S_{2j} = \Sigma_{i=r_1+1}^{r_2} Y_{ij}$ and $S_{3j} = \Sigma_{i=r_2+1}^{13} Y_{ij}$, for $j=1, 2$, and where $\Gamma(\ )$ is the gamma function.

Table 2 contains the joint posterior probabilities for all possible pairs of changepoints for this data set using the prior in expression (5). The joint posterior modal value occurs at $r_1=4$, $r_2=5$.

3. Gibbs Sampler: Motivation and Methodology

The Gibbs sampler is one of a class of Markov chain Monte Carlo algorithms that facilitate practical statistical problem solving. In a Bayesian framework, the common objective is to produce posterior densities for, or estimates of, parameters of interest. For simple models, such as that in example 1, analytic calculation is possible, and results are directly available. For more complex models, such as would arise in the scribes example if the conjugate prior in expression (5) were to be replaced by a non-conjugate form, recourse to numerical integration is necessary. However, for complex models in which, for example, there are non-linear terms in the likelihood or prior, or where the parameter space is constrained in some awkward fashion, or where the number of unknown parameters is large, conventional numerical techniques are often insufficiently accurate, or at least difficult to implement without relevant expertise on the part of the user (see Smith (1991) for a detailed review of analytical and numerical Bayesian techniques). In contrast, the Gibbs sampler is generally straightforward to implement and is easily accessible to the average statistical practitioner, for a wide range of models.

For a general description and illustration of Markov chain Monte Carlo techniques, see, for example, Tierney (1991), Smith and Roberts (1993), Besag and Green (1993) and Gilks *et al.* (1993). For specific explanation and illustrations of the Gibbs sampler algorithm, see, for example, Gelfand and Smith (1990). The algorithm proceeds as follows: given a joint density $p(\boldsymbol{\theta})$ for K-component parameter $\boldsymbol{\theta}$, the set of corresponding 'full conditional' posterior densities derived from the joint density, denoted $p_i(\theta_i \mid \theta_{(i)})$, $i=1, \ldots, K$ (where $\theta_{(i)}$ represents the $(K-1)$-component vector $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K)$), and a vector of initial values for the K components, denoted $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_K^{(0)})$, it is required that we sample in turn from the each of the full conditionals, with the set of conditioning variables initially set equal to $\boldsymbol{\theta}^0$, but with a component of this vector being updated whenever a new value for the corresponding parameter is sampled, i.e. for the first iteration we sample

$$\begin{aligned}
\theta_1^{(1)} &\sim p_1(\theta_1 \mid \theta_2^{(0)}, \ldots, \theta_K^{(0)}),\\
\theta_2^{(1)} &\sim p_2(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_K^{(0)}),\\
&\vdots\\
\theta_{K-1}^{(1)} &\sim p_{K-1}(\theta_{K-1} \mid \theta_1^{(1)}, \ldots, \theta_{K-2}^{(1)}, \theta_K^{(0)}),\\
\theta_K^{(1)} &\sim p_K(\theta_K \mid \theta_1^{(1)}, \ldots, \theta_{K-1}^{(1)})
\end{aligned} \tag{6}$$

to obtain a new vector $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \ldots, \theta_K^{(1)})$. Proceeding thus, when the tth iteration is complete, a new vector of values $\boldsymbol{\theta}^{(t)}$ will have been obtained. Then, subject to certain mild regularity conditions on the joint and conditional densities, it can be shown that the sampled vector tends in distribution to a sample from the

joint density p as t tends to infinity. Also, estimates of posterior densities or quantities may be computed during the iterative procedure, by, for example, formation of ergodic sample averages (expectations) of the relevant conditional functionals obtained at each iteration. To produce, say, a density estimate for the marginal density of parameter θ_1 after T iterations, we might form the average

$$\hat{p}_1(\theta_1) = \frac{1}{T}\sum_{t=1}^{T} p_1(\theta_1 | \theta_2^{(t)}, \ldots, \theta_K^{(t)}),$$

or to produce an estimate of a function g of parameter $\boldsymbol{\theta}$ we would use the average

$$\hat{g}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{\theta}^{(t)}).$$

Alternatively, we might simply wish to use the collection of sampled vectors in some sort of summary display. The Gibbs sampler is usually easy to implement as, although the joint probability structure may be complex, the (essentially univariate) conditional structure is generally much more simple. Constraints are readily incorporated by inspecting their implication for the conditional densities and adjusting the sampling accordingly. The algorithm is also applicable to problems in which some parameters are discrete.

Precise details of implementation of Markov chain Monte Carlo techniques (optimal strategies, estimation, sampling considerations, the implications for parameterization etc.) vary between individual problems and are the subject of much current research. Here, we outline the way that the algorithm would typically be implemented in the changepoint problem.

3.1. *Computation in Changepoint Problem using Gibbs Sampler*

Extending the notation of Section 2.1, consider the sequence of random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ assumed to have k changepoints $(r_1, \ldots, r_k)$ of unknown position but where k is presumed known. Let $r_0 = 0$ and $r_{k+1} = n$. Let the sampling distribution of $\mathbf{Y}_j = (Y_{r_{j-1}+1}, \ldots, Y_{r_j})$ have parameters θ_j and $(Y_{r_{j-1}+1}, \ldots, Y_{r_j})$ be conditionally independent given θ_j, $j = 1, \ldots, k+1$. Finally, let $(\mathbf{Y}_1, \ldots, \mathbf{Y}_{k+1})$ be conditionally independent given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k+1})$.

To implement the Gibbs sampler, we must have explicit forms for the set of full conditional posterior distributions for the unknown parameters. Now, in our formulation, an important simplification in the conditional structure is immediately obvious. From the conditional independence assumptions, it is clear that the conditional posterior distribution of r_j given $(r_0, \ldots, r_{j-1}, r_{j+1}, \ldots, r_{k+1})$ is only dependent on changepoints r_{j-1} and r_{j+1}, and the data between them, namely $\mathbf{Y}_j$ and $\mathbf{Y}_{j+1}$, i.e.

$$[r_j | r_0, \ldots, r_{j-1}, r_{j+1}, \ldots, r_{k+1}, \mathbf{Y}, \boldsymbol{\psi}] \equiv [r_j | r_{j-1}, r_{j+1}, \mathbf{Y}_j, \mathbf{Y}_{j+1}, \boldsymbol{\psi}]. \quad (7)$$

Furthermore, the conditional distribution in expression (7) is precisely the one-changepoint posterior distribution as indicated in expression (2), with the range of r_j being restricted to $r_{j-1}+1, \ldots, r_{j+1}$. We may readily write down the full conditional posterior distributions $[r_j | r_{j-1}, r_{j+1}, \mathbf{Y}, \boldsymbol{\psi}]$ by using standard techni-

ques, for each of the parameters r_j, $j=1, \ldots, k$. Also, as the distributions in expression (7) are univariate and discrete, sampling from them during the Gibbs sampler cycle will be straightforward by simple distribution function inversion. Because of these factors, it appears that discrete multiple-changepoint problems are especially well suited for analysis using the Gibbs sampler. Indeed, if the number of presumed changepoints is large, it will be positively advantageous to use the algorithm, as exact calculations (normalization in particular) will be extremely computationally laborious, requiring summation over all possible ordered k-tuples of changepoints.

In this formulation we have integrated out the parameters of secondary interest, namely $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k+1})$. If the θ_js were of interest, and we wanted to calculate the marginal posterior densities $[\theta_j | \mathbf{Y}, \boldsymbol{\psi}]$, we could extend the Gibbs sampler by including the $k+1$ conditional posterior densities

$$[\theta_j | r_0, \ldots, r_{k+1}, \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_{k+1}, \mathbf{Y}, \boldsymbol{\psi}]$$

in the sampling cycle described above. It is easily seen that this conditional posterior density for θ_j simplifies to

$$[\theta_j | r_{j-1}, r_j, \mathbf{Y}_j, \boldsymbol{\psi}] \propto \prod_{i=r_{j-1}+1}^{r_j} [Y_i | \theta_j][\theta_j | \boldsymbol{\psi}]. \tag{8}$$

Initially, we concentrate solely on the changepoint posterior marginals.

3.2. *Sampling-based Analysis of Scribes Data*

We now attempt to reproduce the analytical results calculated for the scribes data in Section 2.2, by computing the joint posterior distribution of pairs of changepoints using the Gibbs sampler. In the scribes example, under the prior specification in expression (4), it is easily seen that the full conditional posterior distributions for r_1 and r_2 are given by

$$[r_1 | r_2, \mathbf{Y}, \boldsymbol{\psi}] \propto \frac{\Gamma(S_{11}+1)\,\Gamma(S_{12}+1)}{\Gamma(S_{11}+S_{12}+2)} \frac{\Gamma(S_{21}+1)\,\Gamma(S_{22}+1)}{\Gamma(S_{21}+S_{22}+2)} \qquad 1 \leqslant r_1 < r_2 \tag{9}$$

with r_2 fixed and

$$[r_2 | r_1, \mathbf{Y}, \boldsymbol{\psi}] \propto \frac{\Gamma(S_{21}+1)\,\Gamma(S_{22}+1)}{\Gamma(S_{21}+S_{22}+2)} \frac{\Gamma(S_{31}+1)\,\Gamma(S_{32}+1)}{\Gamma(S_{31}+S_{32}+2)} \qquad r_1 < r_2 < 13 \tag{10}$$

with r_1 fixed—note how these conditional distributions can be deduced immediately by inspection of expression (5). Note, also, that if we were to include the three full conditionals for the unknown $(\theta_1, \theta_2, \theta_3)$, they would also be straightforward. Because of the (conjugate) prior specification, it is easy to see that these full conditionals would be of the standard beta form, i.e.

$$[\theta_j | r_1, r_2, \mathbf{Y}_j, \boldsymbol{\psi}] \equiv \text{beta}(S_{j1}+1, S_{j2}+1) \tag{11}$$

for $j=1, 2, 3$.

Table 3 contains the results obtained after 300 iterations of the Gibbs sampler. The contents represent relative frequencies if the occurrence of all possible pairs of changepoint values, averaged over 1000 independent replications. Clearly, there

TABLE 3
Example 1—approximate joint posterior distribution computed via the Gibbs sampler

r_2	*Probabilities for the following values of* r_1*:*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
2	0.000	–	–	–	–	–	–	–	–	–	–
3	0.001	0.000	–	–	–	–	–	–	–	–	–
4	0.001	0.000	0.000	–	–	–	–	–	–	–	–
5	0.063	0.028	0.040	0.337	–	–	–	–	–	–	–
6	0.061	0.017	0.027	0.040	0.052	–	–	–	–	–	–
7	0.013	0.004	0.003	0.002	0.036	0.017	–	–	–	–	–
8	0.006	0.000	0.004	0.000	0.028	0.017	0.003	–	–	–	–
9	0.000	0.000	0.000	0.000	0.020	0.011	0.003	0.001	–	–	–
10	0.001	0.001	0.000	0.000	0.020	0.016	0.000	0.000	0.000	–	–
11	0.001	0.000	0.000	0.000	0.025	0.021	0.002	0.003	0.001	0.000	–
12	0.000	0.000	0.001	0.000	0.039	0.024	0.008	0.002	0.000	0.000	0.000

is close agreement between the exact and approximate joint distributions in this case. Also, on further experimentation, it became clear that the Gibbs sampler could approximately reproduce the exact joint distribution in many fewer than 300 iterations. Inspection of the 1000 replicate pairs after as little as 10 or 20 iterations confirms this.

4. Changepoint Identification—Continuous Case

In Section 2, we described the usual discrete changepoint formulation. We now consider the continuous analogue. As noted earlier, in many practical examples, it is realistic to presume that the data model, although possibly undergoing functional change, is continuous in time, more specifically, continuous at the changepoints. In this case, we might regard the changepoint parameter as continuous, as the discrete observation ordinates will not usually coincide with the changepoint ordinate.

Consider, for example, a simple continuous one-changepoint model, the switching straight line regression model (see, for example, Smith and Cook (1980)). Here, the model may be written

$$\begin{aligned} Y_t &= a_1 + b_1 t + e_t \qquad t \leqslant \gamma, \\ Y_t &= a_2 + b_2 t + e_t \qquad t > \gamma \end{aligned} \tag{12}$$

where the e_t are independent and identically distributed normal 'error' variates with variance σ^2, and where the constraint $a_1 + b_1\gamma = a_2 + b_2\gamma$ ensures continuity. The six-parameter model (12) is overparameterized (by one parameter) because of this equality constraint.

It is important here to note that a discrete changepoint representation of model (12) is not possible. To see this, consider the discretized version of model (12), where the changepoint r is such that $t_r \leqslant \gamma < t_{r+1}$. The model still has five parameters owing to the continuity constraint, but any model involving γ and r is meaningless, as r, being the index of the largest t-value not greater than γ, is a deterministic function of γ. So consider the model parameterized by, for instance,

r, a_1, b_1, a_2 and σ. In this case, the second regression line is not well defined $-b_2$ cannot be uniquely expressed as a function of these parameters. This is true for any parameterization involving r rather than γ.

Thus, in the continuous example, we are bound to use a continuous changepoint representation. Similarly, in the non-continuous case, we are bound to use a discrete changepoint parameterization, as then the position of γ, the ordinate of the intercept of the two lines, is irrelevant, as it does not in any way relate to the position at which the data model changes. More precisely, in the non-continuous case Y is independent of γ given r, whereas in the continuous case Y is independent of r given γ.

4.1. *Example 2—Switching Linear Regression Model*

Given that we must treat the changepoint parameter in the switching straight line model as continuous, the inference problem immediately becomes analytically intractable, and also awkward to deal with by conventional techniques. In the Gibbs sampler approach, however, many of the difficulties can be avoided because of the relatively simple conditional structure.

For model (12), the likelihood in the a_1, b_1, a_2, γ and σ parameterization, with $b_2 = b_1 + (a_1 - a_2)/\gamma$, takes the form

$$\frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\left[\sum_{t\leqslant\gamma}(Y_t - a_1 - b_1 t)^2 + \sum_{t>\gamma}\left\{Y_t - a_2 - b_1 t - \frac{(a_1 - a_2)t}{\gamma}\right\}^2\right]\right), \quad (13)$$

with a_1, b_1 and a_2 unconstrained, and γ constrained to lie between the first and last ordinates of the observed data. As the continuous parameter γ defines the ranges of summation of the exponent sums of squares in expression (13), the likelihood surface is ill suited to analysis via conventional numerical methods, being possibly only piecewise continuous, with discontinuities at the data ordinates. Also, the likelihood is non-linear in γ, and so analytic approximation (via modal curvature techniques at least) might seem inappropriate.

In a Gibbs sampler context, however, these immediate difficulties disappear. By inspection of the functional form of expression (13), it is relatively easy to see that, under a non-informative prior (uniform) specification for each of the parameters, the full conditional posterior distributions for a_1, b_1 and a_2 are normal, and for σ^2 inverse gamma. The conditional posterior density of γ is of non-standard form. To sample from it, we use rejection sampling in the way proposed by Smith and Gelfand (1992), i.e. we sample x from the conditional prior and accept x as an observation from the conditional posterior if $uM/f(x) \leqslant 1$, where u is a $U(0, 1)$ variate, $f(\ )$ is the conditional likelihood function given by expression (13) and M is the maximum value of the conditional likelihood. However, as the conditional posterior density for the changepoint parameter is possibly only piecewise continuous (discontinuous at the data points), we actually sample from a piecewise uniform density constructed as follows: we maximize the conditional density on each between-datum segment and then consider a uniform density for that interval, with level proportional to the maximum value. After normalization, we have a piecewise uniform density that dominates the changepoint conditional from which we can obtain a sample straightforwardly. Rejection then proceeds in the usual

fashion. This technique is considerably more efficient than rejection sampling using a single uniform density.

4.1.1. *Analysis of switching regression model data*

Table 4 contains the data analysed by Smith and Cook (1980), relating to two patients who have each undergone a failed kidney transplant, where the observed response is the level of a substance reflecting the healthy or otherwise functioning

TABLE 4
Example 2—renal transplant data

Patient	*Values of Y_t for the following values of t:*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	48.6	58.0	62.3	71.9	80.6	54.8	49.3	43.4		
B	36.8	46.5	50.8	66.2	75.0	71.5	68.5	60.5	31.5	19.0

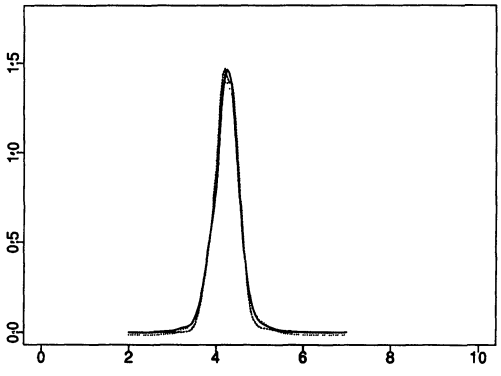

Fig. 4. γ (average conditionals) densities for patient A: --------, after 500 iterations; ·····, after 1000 iterations; ———, after 2000 iterations

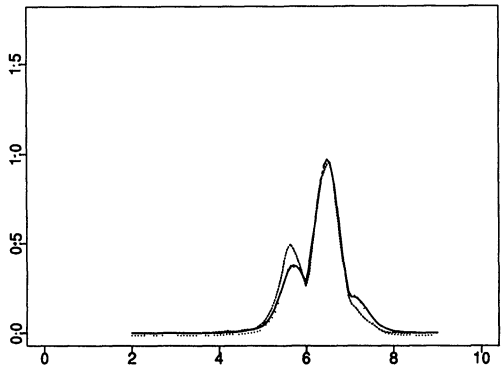

Fig. 5. γ (average conditionals) densities for patient B: --------, after 500 iterations; ·····, after 1000 iterations; ———, after 2000 iterations

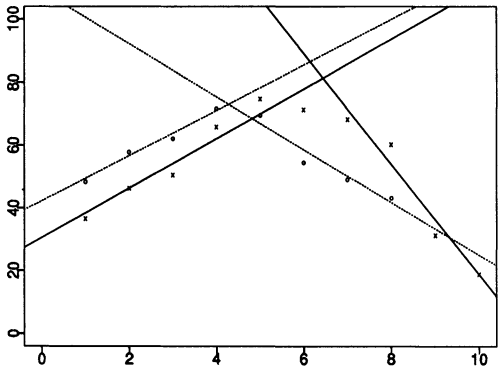

Fig. 6. Data and estimated lines: ○, --------, patient A; ×, ———, patient B

TABLE 5
Example 2—regression line parameter estimates†

Patient	a_1	b_1	a_2	b_2
A	42.30 (3.47)	7.20 (1.33)	109.00 (8.17)	−8.40 (1.25)
B	30.58 (6.81)	7.88 (2.01)	194.44 (33.95)	−17.56 (3.84)

†Standard error are given in parentheses.

of the transplanted kidney. The changepoint represents the point in time at which the kidney is rejected. The data model is presumed to be that in equation (12), the model is presumed to be continuous in time and all the parameters are unknown.

Using the Gibbs sampler, approximate posterior densities for the changepoint positions and other parameters for both patients can be computed. Figs 4 and 5 depict the posterior densities for the changepoint positions for patients A and B respectively. The densities were computed by using ergodic averaging of the relevant full conditional density as described in Section 3 after 500 (broken curve), 1000 (dotted curve) and 2000 (full curve) iterations. In both cases, satisfactory convergence appears to have occurred before 2000 iterations. Fig. 6 depicts the data and 'estimated' regression lines for each patient, and the sample-based parameter estimates are presented in Table 5. The multimodality of the estimated marginal posterior density for γ for patient B is reflected in the inflated sample standard errors for the parameters of the regression lines. Note that, in the first regime, i.e. before the changepoint and rejection of the kidney, the estimates of the slope parameters are approximately equal.

4.2. *Example 3—Multiphase Growth Model*

We now study the continuous version of the multiple-changepoint problem via the example in Schulze (1984) relating to the growth in mass of bacteria in a nutritive

medium. It is believed that the observed mass Y varies with time t according to the model

$$Y_t = f(\boldsymbol{\theta}, \boldsymbol{\gamma}, t) + \epsilon_t \tag{14}$$

where

$$f(\boldsymbol{\theta}, \boldsymbol{\gamma}, t) = \begin{cases} a_1 \exp(b_1 t) & t \leqslant \gamma_1, \\ c_1 & \gamma_1 < t \leqslant \gamma_2, \\ a_2 \exp(b_2 t) & \gamma_2 < t \leqslant \gamma_3, \\ c_2 & \gamma_3 < t, \end{cases}$$

the ϵ_t are independent and identically distributed $N(0, \sigma^2)$, $\boldsymbol{\theta} = (a_1, b_1, c_1, a_2, b_2, c_2)$ are the unknown growth curve parameters and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ are the unknown changepoints, here interpreted as being continuous parameters, corresponding to the growth phase endings. We observe Y at specific values of t, none of which need coincide with any of the γ_i but belong to one of the four regimes, in accordance with the values of the γ_i. For the growth curve to be continuous, we must have that

$$\begin{gathered} a_1 \exp(b_1 \gamma_1) = c_1 = a_2 \exp(b_2 \gamma_2), \\ a_2 \exp(b_2 \gamma_3) = c_2. \end{gathered} \tag{15}$$

As before, the model in this form is overparameterized because of the constraints imposed by the continuity assumption. We also impose other natural constraints, namely that all parameters are strictly positive, and that $\gamma_1 < \gamma_2 < \gamma_3$. The likelihood in its overparameterized form is proportional to

$$\frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{t \leqslant \gamma_1} \{Y_t - a_1 \exp(b_1 t)\}^2 + \sum_{\gamma_1 < t \leqslant \gamma_2} (Y_t - c_1)^2 \right.\right.$$
$$\left.\left. + \sum_{\gamma_2 < t \leqslant \gamma_3} \{Y_t - a_2 \exp(b_2 t)\}^2 + \sum_{\gamma_3 < t} (Y_t - c_2)^2 \right] \right) \tag{16}$$

with three of the nine parameters being explicit functions of the other six. This problem is analytically intractable whatever the choice of prior, and, once again, owing to the parameter constraints, discontinuities in the likelihood and extreme non-linearity, it is also extremely difficult to solve by conventional numerical techniques.

4.3. *Gibbs Sampler for Growth Curve Problem*

For solution by using the Gibbs sampler, we first need to obtain explicit forms for the full conditional posterior distributions. First, we must select a suitable parameterization. Initially, it seems advantageous to parameterize in terms of $\boldsymbol{\theta}$ and to eliminate $\boldsymbol{\gamma}$ by using expression (15), despite the fact that we are primarily interested in the changepoint positions, as this appears to lead to standard form (normal) conditional distributions for four of the six parameters. However, as an allocation of points to regimes depends on $\boldsymbol{\gamma}$, and hence on $\boldsymbol{\theta}$, these conditional distributions are only piecewise normal. This, in conjunction with the nature of the constraints when interpreted via $\boldsymbol{\theta}$ as in equations (15), leads us to seek another parameterization.

Despite the form of their conditional densities, it is desirable to parameterize in terms of $\boldsymbol{\gamma}$, as the constraint $0 < \gamma_1 < \gamma_2 < \gamma_3 < T_{\max}$ leads to finite ranges for these parameters. This is of use when designing rejection sampling routines. Also, from the inversion formulae in equations (15), it seems more straightforward to retain b_1 and b_2, to avoid the use of computationally expensive logarithm operations. Thus we choose to parameterize in terms of $\boldsymbol{\gamma}$, b_1, b_2 and c_1, and eliminate a_1, a_2 and c_2 from expression (16), i.e. setting

$$\left.\begin{aligned} a_1 &= c_1 \exp(-b_1\gamma_1), \\ a_2 &= c_1 \exp(-b_2\gamma_2), \\ c_2 &= c_1 \exp\{b_2(\gamma_3-\gamma_2)\}. \end{aligned}\right\} \tag{17}$$

Recall that, ultimately, we shall be able to obtain a sample of a_1, a_2 and c_2 as a function of the samples of $\boldsymbol{\gamma}$, b_1, b_2 and c_1 simply by mapping via the functions in equations (17).

The likelihood in the chosen parameterization is proportional to $\sigma^{-n} \times \exp(-\mathrm{SSQ}/2\sigma^2)$, where

$$\begin{aligned} \mathrm{SSQ} = \sum_{t \leqslant \gamma_1} [Y_t - c_1 \exp\{b_1(t-\gamma_1)\}]^2 + \sum_{\gamma_1 < t \leqslant \gamma_2} (Y_t - c_1)^2 + \sum_{\gamma_2 < t \leqslant \gamma_3} [Y_t \\ - c_1 \exp\{b_2(t-\gamma_2)\}]^2 + \sum_{\gamma_3 < t} [Y_t - c_1 \exp\{b_2(\gamma_3-\gamma_2)\}]^2. \end{aligned} \tag{18}$$

The relative complexity of equation (18) illustrates the potential difficulties that would be encountered when using conventional numerical techniques. Two other factors must be noted. First, we must have at least one observation per regime for identifiability, so that, in terms of the changepoint parameters,

$$\left.\begin{aligned} T(1) &\leqslant \gamma_1 < T(G_2), \\ T(G_1+1) &\leqslant \gamma_2 < T(G_3), \\ T(G_2+1) &\leqslant \gamma_3 < T(n), \end{aligned}\right\} \tag{19}$$

where G_i is the index of the largest of the observed t-values not greater than γ_i, when the ts are ranked in ascending order in the vector $T(\ )$, for $i = 1, 2, 3$. Secondly, it is clear from equation (18) that with a non-informative uniform prior for b_1 the joint posterior density and conditional posterior density for b_1 will be improper, so a proper prior for b_1 must be specified.

For illustration, we shall use independent exponential priors for b_1 and b_2, and a non-informative uniform prior for c_1, in conjunction with the uniform joint or conditional priors for $\boldsymbol{\gamma}$ indicated by inequalities (19). Thus, the full conditional posterior densities for the parameters are given directly by equation (18), with the domains for the growth curve parameter densities as defined by inequalities (19), and b_1, b_2 and $c_1 > 0$.

The conditional posterior densities as indicated by equation (18) are of relatively simple forms, being exponentials of sum(s) of squares. The conditional density of c_1 is normal constrained to the positive real half-line. A one-for-one sampling sampling technique to sample from this constrained normal density is approximate cumulative density function inversion, i.e. if u is $U(0, 1)$ distributed

then an observation x from a unit normal density restricted to the positive real half-line may be obtained as the solution to

$$\Phi(x) = 0.5 + 0.5u,$$

where $\Phi(\ )$ is the unit normal cumulative density function. The appropriate value of x can be found numerically.

The conditional posterior densities of γ_1, γ_2, γ_3, b_1 and b_2 are of non-standard form. To sample from them, we use the rejection sampling routine described above, sampling initially from a conditional (uniform or exponential) prior, and rejecting using the conditional likelihood function. Again, the argument giving the maximum of the likelihood equates precisely with the conditional least squares estimate.

TABLE 6
Example 3—growth curve data

t	Y_t	t	Y_t
0.00	0.0226	2.50	0.0862
0.25	0.0250	2.75	0.1021
0.50	0.0300	3.00	0.1267
0.75	0.0472	3.25	0.1264
1.00	0.0623	3.50	0.1416
1.25	0.0661	3.75	0.1468
1.50	0.0770	4.00	0.1535
1.75	0.0772	4.25	0.1491
2.00	0.0804	4.50	0.1517
2.25	0.0868		

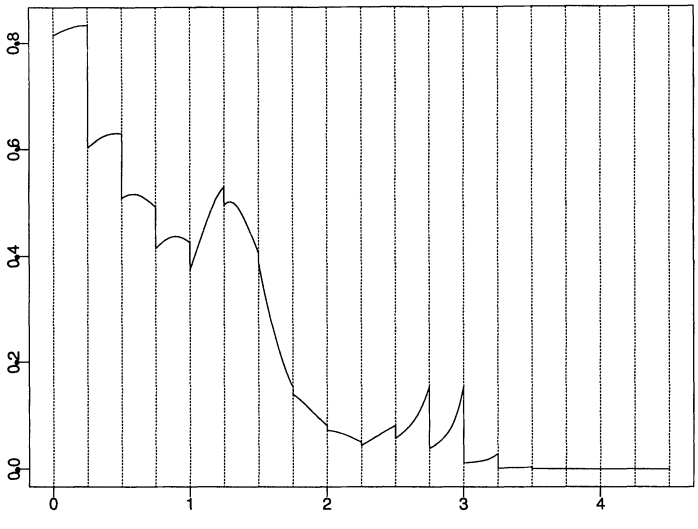

Fig. 7. Estimated marginal posterior density of γ_1

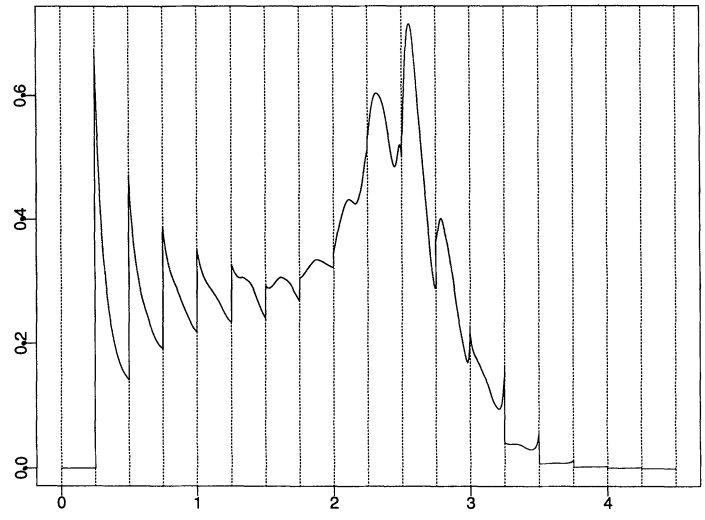

Fig. 8. Estimated marginal posterior density of γ_2

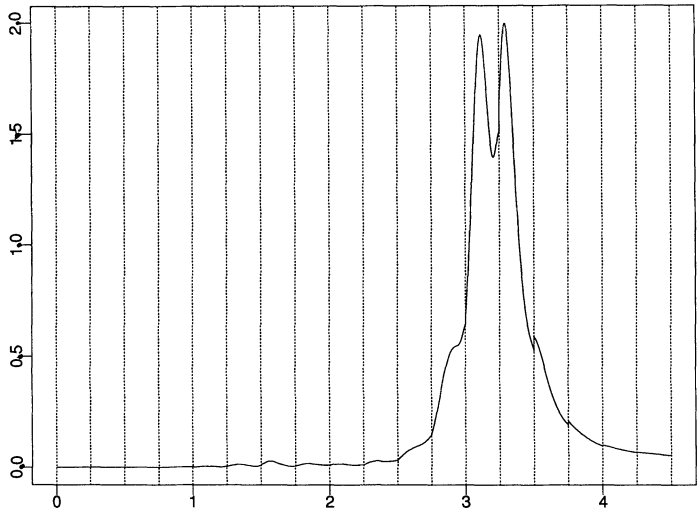

Fig. 9. Estimated marginal posterior density of γ_3

4.4. *Analysis of Growth Curve Data*

The (simulated) data reproduced from Schulze (1984) are presented in Table 6. The starting values for the Gibbs sampler are obtained by simple *ad hoc* methods from the data. The prior specifications for b_1 and b_2 are independent exponential(1.0) priors, for illustration.

Figs 7, 8 and 9 are density estimates calculated by pointwise averaging of the successive conditional densities, as used in the Gibbs sampler, for γ_1, γ_2 and γ_3 respectively. The broken vertical lines indicate the position of the data point ordinates, and there is clear evidence that the marginal densities are not continuous

TABLE 7
Example 3—approximate values of $p_j = Pr\{T(j) < \gamma_i < T(j+1)\}$

j	*Values of* p_j *for the following values of* γ_i: γ_1	γ_2	γ_3
1	0.208	0.000	0.000
2	0.156	0.070	0.000
3	0.128	0.067	0.000
4	0.109	0.069	0.000
5	0.114	0.070	0.001
6	0.115	0.071	0.002
7	0.063	0.075	0.003
8	0.028	0.082	0.004
9	0.016	0.106	0.005
10	0.016	0.137	0.007
11	0.025	0.125	0.024
12	0.018	0.078	0.105
13	0.004	0.036	0.375
14	0.001	0.009	0.310
15	0.001	0.002	0.090
16	0.000	0.001	0.036
17	0.000	0.000	0.021
18	0.000	0.000	0.015

at the data points. It is possible to approximate quantities such as $p_j = \Pr\{T(j) < \gamma_i < T(j+1)\}$ from these density estimates, possibly by trapezium-rule-type techniques. These values are recorded in Table 7 and can be regarded as roughly equivalent to the discrete posterior changepoint distribution described in Sections 1 and 2 (although, as noted earlier, discrete changepoint marginal distributions are not directly available here). The accuracy of such estimates may be investigated by means of replicate Gibbs sampler implementations.

Examination of the time series plots of the output streams of sampled changepoint variates reveals a high serial correlation. This is intuitively reasonable (given the parameterization), but unfortunate, as it is widely agreed that high parameter dependence lessens the rate of convergence of the Gibbs sampler. The results given here were obtained after a fairly large number of iterations (200000) and were confirmed by inspection of similar statistics after even longer runs. However, the sample statistics etc. were also reasonably stable after shorter runs (10000). It was noticeable that the 'discrete' parameter changepoint probabilities, as given in Table 7, stabilized much more quickly than did the density plots. This indicates, as in example 1, that estimates of certain of our quantities will be adequate even though the runs involved are relatively short. In turn, this implies that any assessment of convergence of the Gibbs sampler must in part relate to the stability of these estimates.

5. Discussion

We have shown how the Gibbs sampler can be used to aid inference in simple and more complex single- and multiple-changepoint models. The algorithm itself

is relatively straightforward to implement—we merely write down the functional form of the unnormalized joint posterior density, deduce the forms of the full conditional by inspection, check that they are proper and then sample from them by using the iterative process until convergence is diagnosed. It also presents the results in an accessible form. In the examples studied, its efficiency varies from high (for the discrete example) to moderate (in the two continuous examples). For instance, the maximization and rejection sampling routines required for the non-standard conditional densities demand most of the computation time. It should be possible to design more specific routines that would improve on these figures. In the third example, where the dependence between parameters is high, the choice of parameterization becomes important. It is doubtful whether any simple linear reparameterization would be of use overall, as it is the very nature of the conditional structure in the parameterization indicated by expressions (13) and (18) (explicitly in terms of the changepoint parameters, with the natural constraints on them) that is the very motivation for using the Gibbs sampler.

Acknowledgements

Part of this work formed the basis of a section of the author's doctoral thesis, carried out at the University of Nottingham, UK. The author would like to thank Adrian Smith and Cliff Litton, and the referees, for their constructive remarks. The work was financially supported by the Science and Engineering Research Council's Complex Stochastic Systems Initiative.

References

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computing. *J. R. Statist. Soc.* B, **55**, 25–37.

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

Broemeling, L. D. (1974) Bayesian inferences about a changing sequence of random variables. *Communs Statist. Theory Meth.*, **3**, 243–255.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992) Hierarchical Bayesian analysis of change-point problems. *Appl. Statist.*, **41**, 389–405.

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

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

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993) Modelling complexity: applications of Gibbs sampling in medicine. *J. R. Statist. Soc.* B, **55**, 39–52.

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

——— (1971) Inference about the change-point from cumulative sum tests. *Biometrika*, **58**, 509–523.

Pettitt, A. N. (1980) A simple cumulative sum type statistic for the changepoint problem with zero–one observations. *Biometrika*, **67**, 79–84.

Schulze, U. (1984) A method of estimation of changepoints in a multiphase growth model. *Biometr. J.*, **5**, 495–504.

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

——— (1980) Change-point problems: approaches and applications. In *Bayesian Statistics* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 83–89. Valencia: Valencia University Press.

——— (1991) Bayesian computational methods. *Phil. Trans. R. Soc. Lond.* A, **337**, 369–386.
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.
Smith, A. F. M. and Gelfand, A. E. (1992) Bayesian statistics without tears: a sampling–resampling perspective. *Am. Statistn*, **46**, 84–88.
Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.
Tierney, L. (1991) Markov chains for exploring posterior distributions. *Technical Report 560*. School of Statistics, University of Minnesota, Minneapolis.