A Bayesian Analysis for Change Point Problems

DANIEL BARRY and J. A. HARTIGAN*

A sequence of observations undergoes sudden changes at unknown times. We model the process by supposing that there is an underlying sequence of parameters partitioned into contiguous blocks of equal parameter values; the beginning of each block is said to be a change point. Observations are then assumed to be independent in different blocks given the sequence of parameters. In a Bayesian analysis it is necessary to give probability distributions to both the change points and the parameters. We use product partition models (Barry and Hartigan 1992), which assume that the probability of any partition is proportional to a product of prior cohesions, one for each block in the partition, and that given the blocks the parameters in different blocks have independent prior distributions. Given the observations a new product partition model holds, with posterior cohesions for the blocks and new independent block posterior distributions for parameters. The product model thus provides a convenient machinery for allowing the data to weight the partitions likely to hold; inference about particular parameters may then be made by first conditioning on the partition, and then averaging over all partitions. The parameter values may be estimated exactly in $O(n^3)$ calculations, or to an adequate approximation by Markov sampling techniques that are $O(n)$ in the number of observations. The Markov sampling computations are thus practicable for long sequences. We compare this model with a number of alternative approaches to fitting change points and parameters when the error distribution is normal, then show that the proposed method is superior to the alternatives in detecting sharp short-lived changes in the parameters.

KEY WORDS: Change points; Product partition models.

1. INTRODUCTION

Lombard (1987) considered a sequence of data consisting of the radii of 100 circular indentations cut by a milling machine. (The data are given in our Figure 5.) Lombard suggested that "there may have been two increases in mean, or even a smooth increase, between observations 20 and 40 followed by a decrease at observation 76." We will model such a process as a sequence of observations $X_1, X_2, \ldots, X_n$ ordered in time, each observation X_i being independent with a density depending on a parameter $\theta_i \in \Theta$ whose value may change from one observation to the next. In this article we examine problems of inference when there exists an unknown partition ρ of the set $\{1, 2, \ldots, n\}$ into contiguous sets or *blocks* such that the sequence $\theta_1, \theta_2, \ldots, \theta_n$ is constant within blocks; that is, there exists a partition $\rho = (i_0, i_1, \ldots, i_b)$ of the set $\{1, 2, \ldots, n\}$ such that

$$0 = i_0 < i_1 < i_2 < \cdots < i_b = n$$

and

$$\theta_i = \theta_{i_r} \qquad i_{r-1} < i \le i_r$$

for $r = 1, 2, \ldots, b$. The parameter values change at the change points $i_1 + 1, i_2 + 1, \ldots, i_{b-1} + 1$. It will be convenient to identify the sequence of time points $i + 1, \ldots, j$ by the symbol ij and the corresponding observations $X_{i+1}, \ldots, X_j$ by X_{ij}.

Given the partition and given the parameters, the observations $X_1, \ldots, X_n$ are assumed to be independent in different blocks, having density $\prod_{j=1}^{b} f_{i_{j-1}i_j}(X_{i_{j-1}i_j})|\theta_j)$, where $f_{ij}(x_{ij}|\theta_j)$ denotes the density of x_{ij} given $\theta_{i+1} = \theta_{i+2} = \cdots = \theta_j$. (The notation $f(\cdot)$ will be used for densities, and $f(\cdot\,|\,\cdot)$ for conditional densities, of all sorts.)

In the product partition model (Barry and Hartigan 1990), the partition is randomly selected according to a *product partition* distribution: The probability of a partition $\rho = (i_0, i_1, i_2, \ldots, i_b = n)$ is

$$f(\rho) = Kc_{i_0i_1}c_{i_1i_2}\cdots c_{i_{b-1}i_b},$$

determined by prior *cohesions* c_{ij} specified for each possible block ij. As a consequence of the model, the number of blocks b is a random variable ranging from 1 to n. See Hartigan (1990) for product partition models for general partitions.

We now construct a prior distribution for $\theta_1, \theta_2, \ldots, \theta_n$ as follows. Given a partition ρ with b blocks, $\theta_{i_1}, \theta_{i_2}, \ldots, \theta_{i_b}$ are independent, with θ_{i_j} having density $f_{i_{j-1}i_j}(\theta_{i_j})$ with respect to some measure on Θ (which we will represent in integrals as $d\theta$). The joint distribution of all parameters is now determined, because all parameters in the jth block are equal to θ_{i_j}. The density $f_{ij}(\theta_j)$ will be called the *block prior density*. The combined model for partition and parameters is the product partition model.

For a given block ij, the density of the observations X_{ij} is

$$f_{ij}(X_{ij}) = \int f_{ij}(X_{ij}|\theta)f_{ij}(\theta)\,d\theta.$$

The conditional distribution of partition and parameters given the observations is also a product partition model with posterior cohesions $c_{ij}f_{ij}(X_{ij})$ and block posterior density $f_{ij}(\theta_j|X_{ij}) = f_{ij}(\theta_j)f_{ij}(X_{ij}|\theta_j)]/f_{ij}(X_{ij})$.

Thus product models on partitions give us a workable framework for making inferences about change points based on the data $X_1, \ldots, X_n$. Even if the initial probability model for partitions and parameters is not a product model, in circumstances where the observations are sharp enough to dominate the prior distribution the posterior distribution for partitions will be usefully approximated by a product model, because the contribution from the density of observations given parameters will be in product form.

Many different types of problems may be tackled in this framework. We will consider in some detail the normal errors

* Daniel Barry is Lecturer in Statistics, University College, Cork, Ireland; J. A. Hartigan is Professor of Statistics, Yale University, New Haven, CT 06520. This research was partially supported by National Science Foundation Grant DMS-8617919.

model, in which it is supposed that the observations X_i are independent $N(\mu_i, \sigma^2)$. The normal assumption could be replaced by any other parametric assumption and similar analyses carried through. Also, the independence hypothesis could be weakened, because all that is required is that, given the partition and given the parameters, observations in different blocks are mutually independent.

As an alternative to the product partition model, we might suppose that the parameter sequence θ_i forms a Markov chain; for there to be constant blocks, we need a transition distribution in the following form: Given θ_i, θ_{i+1} equals θ_i with probability $1 - p_i$ or has density $f(\theta_{i+1} | \theta_i)$ with probability p_i.

If the p_i are small, then the blocks of constant parameter values will be long. If the conditional density of θ_{i+1} given θ_i, when there is a change, does not depend on θ_i, then the Markov model is a product partition model. In general, however, the two models are different.

We will consider a number of change point analyses for the "normal errors" model. One kind of analysis uses penalized likelihood to select the partition, without any further probabilistic assumptions about the partition or the parameters. For example, Yao (1988) selected ρ to maximize the Schwarz criterion

$$L(\rho) = -n \log \hat{\sigma}_\rho - b \log n,$$

where $\hat{\sigma}_\rho$ is the maximum likelihood estimate of σ for partition ρ and b is the number of blocks in ρ.

There are various Bayesian analyses depending on the particular prior distributions chosen for partitions and parameters. One such model is due to Chernoff and Zacks (1964). The sequence of parameter values forms a Markov chain in which, given μ_i, $\mu_{i+1} = \mu_i$ with probability $1 - p$ and is distributed as $N(\mu_i, \sigma_0^2)$ with probability p. Thus there is a change at point i with probability p, and if the change occurs, the parameter value changes by a normally distributed increment. This model is a Markov model, but not a product partition model.

Yao (1984) gave the same probability to a change; but if the change occurs, the new parameter value is distributed as $N(\mu_0, \sigma_0^2)$. Thus parameter values in different blocks are independently distributed. This model is both a Markov model and a product partition model. Because it is a product partition model, it is possible to compute exact posterior means for the μ_i in $O(n^3)$ computations. There remain difficult problems in estimating the parameters p, σ, σ_0, and μ_0; Yao suggested using maximum likelihood with the constraint that $p \leq .2$, but in Yao (1984) the computations are done using an approximation for the posterior means rather than strict likelihood calculations.

In Barry and Hartigan (1990) we considered a model in which the partition distribution used cohesions of the form

$$c_{ij} = (j - i)^{-3} \quad \text{for} \quad 0 < i < j < n,$$

$$c_{ij} = (j - i)^{-2} \quad \text{for} \quad i = 0 \quad \text{or} \quad j = n,$$

and

$$c_{0n} = n^{-1}.$$

These partition distributions lead to desirable consistency properties for the posterior partition distribution; for example, if there is no change point, the posterior probability that there is no change point approaches 1. But we did not find that this partition distribution gave better mean square errors for the posterior means than did the Yao model when the parameter p was appropriately estimated from the data.

Here we examine a product partition model in which the probability of change at point i is p, independently at each point i. The prior distribution of the parameter μ_j for block ij is $N(\mu_0, \sigma_0^2/(j - i))$. In asserting this prior, we are expecting larger deviations from μ_0 in short blocks than in long blocks; it is not feasible to identify small deviations in short blocks, a presumption that is built into the prior probabilities. A happy side effect is that the posterior distributions of partitions and parameters are much simplified, and it becomes possible to explicitly integrate out the nuisance parameters p, σ, σ_0, and μ_0, thus avoiding the un-Bayesian inelegancies of maximum likelihood on a likelihood function that may have several modes.

We compare methods of Schwarz (given in Yao 1988), Chernoff and Zacks (1964), Yao (1984), and Barry and Hartigan (1990) in a simulation experiment. A sequence of 60 observations is taken from a change point scene, and estimates of the parameters are computed for the four methods; these estimates are maximum likelihood for the optimal choice of ρ using the Schwarz criterion, and posterior means in the other cases. The estimates are exact for Schwarz and for Yao, and based on simulations of posterior distributions for Chernoff and Zacks and Barry and Hartigan. We consider 100 repetitions for each scene, and 20 different scenes.

The Schwarz method is inferior for almost all scenes. The Barry and Hartigan method does best for scenes with outliers or short sharp signals, as might be expected from the prior assumptions. The Yao method does best when the signals are strong and the block lengths are regular. The Chernoff and Zacks method does well when the changes are small and persistent but poorly when the changes are large or persist for only a few observations. Looking over all the scenes, Barry and Hartigan never does much worse than the best method and avoids the disasters of the other methods for sharp short-lived changes.

In Section 7 we apply these four methods to Lombard's data. The data were collected to measure the variability in output from the machine, and any undetected shifts in mean would result in inflated estimates of that variability. Inspecting the fitted means gives an indication of the magnitudes and locations of such mean shifts. Other applications are discussed in Section 8.

What advice do we have for the practitioner who wishes to fit a change point model? First, we need the parameter values in different blocks to be independent. For many kinds of data, if a change occurs, the new parameter values are not too far from the old; for these kinds of data, the Chernoff–Zacks method is more appropriate. Similarly, these techniques will not be effective for spline fitting with, say, piecewise linear models holding in different blocks, because the requirement of independence in different blocks will prevent continuity across segment boundaries. If the independence

assumption is acceptable, then the method we propose can detect change points accurately and efficiently, identify sufficiently large outliers and remove their effect from the rest of the model fitting, and avoid seizing too quickly on modest, briefly sustained random departures as evidence of a new segment. The product partition model is sufficiently simple such that a Bayesian analysis is possible for all parameters; one by-product of the analysis is a posterior distribution for various times being change points and for the number of change points, so that a judgment of whether or not a change point model is necessary can be made after the analysis.

An important practical consideration is that some prior constraint is necessary on the number of change points and on the amount of variability between blocks that we wish to see before constructing new blocks. These are parameters about which the data cannot be relied on to be conclusive. We have found, for example, that constraining the probability of a change at any point to be less than .2, as was first suggested by Yao (1984), is a useful rule. In practical problems we expect to estimate p to be quite a bit less than this, because we do not want to see a change every five observations. An associated rule is that the "signal" variance σ_0^2 should exceed the error variance σ^2 by a factor of 4; this means, for example, that we do not identify an observation as an outlier unless it is at least two standard deviations from the neighbouring segment means.

2. EXACT COMPUTATIONAL PROCEDURES

Although there are 2^{n-1} partitions of n points into blocks of consecutive segments, the product partition model permits calculations of necessary quantities in polynomial time depending on the number of possible blocks, $\binom{n+1}{2}$. Similar recursive calculations are possible for more general product partition models.

Define $\lambda_{ij} = \sum \prod_{k=1}^{b} c_{i_{k-1}i_k}$, where the summation is over all sets of integers $i = i_0 < i_1 < \cdots < i_b = j$. The quantity λ_{ij} is the sum of products of cohesions over all possible partitions of the set $\{i+1, i+2, \ldots, j\}$. Let the relevance r_{ij} be the probability that the block ij is included in the partition ρ. Then

$$r_{ij} = \frac{\lambda_{0i} c_{ij} \lambda_{jn}}{\lambda_{0n}}.$$

The quantities λ_{0i} and λ_{jn} may be calculated in $O(n^2)$ steps using the recursive formulas

$$\lambda_{01} = c_{01},$$

$$\lambda_{0i+1} = c_{0i+1} + \sum_{k=1}^{i} \lambda_{0k} c_{ki+1},$$

$$\lambda_{n-1n} = c_{n-1n},$$

and

$$\lambda_{jn} = c_{jn} + \sum_{k=j}^{n-1} c_{jk} \lambda_{kn}$$

Suppose, for example, that the parameter θ_k is real valued and we wish to compute its posterior expectation given the observations $\mathbf{X}$. The posterior relevances $r_{ij}(\mathbf{X})$ are computed from posterior cohesions by recursive formulas like those just listed. Then

$$E(\theta_k|\mathbf{X}) = \sum_{i<k\le j} E_{ij}(\theta_k|X_{ij}) r_{ij}(\mathbf{X}),$$

where $E_{ij}(\theta_k|X_{ij})$ denotes the posterior expectation of θ_k when the block ij lies in the partition.

For each k there are $O(n^2)$ sets ij that contain k, and so $\{E(\theta_i|\mathbf{X}): i = 1, 2, \ldots, n\}$ may be calculated in $O(n^3)$ steps. The preceding recursive formulas were given in the normal case by Yao (1984).

Because for $\rho = (i_0, i_1, \ldots, i_b)$ the likelihood of $\mathbf{X}$ and ρ is

$$L(\mathbf{X}, \rho) = K \prod_{r=1}^{b} \{f_{i_{r-1}i_r}(\mathbf{X}) c_{i_{r-1}i_r}\},$$

the likelihood of $\mathbf{X}$

$$L(\mathbf{X}) = \sum_{\rho} L(\mathbf{X}, \rho)$$

and its derivatives may also be calculated in $O(n^2)$ steps using similar recursive formulas. These computations are useful in estimating the various nuisance parameters in the model.

3. NORMAL ERRORS

In this section we consider in detail the case where the observations $X_1, X_2, \ldots, X_n$ are independent given the sequence of parameters μ_i with $X_i \sim N(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, n$. To specify a product partition model, we need to choose the prior cohesions c_{ij} and the block prior densities $f_{ij}(\mu)$.

3.1 Choosing the Prior Cohesions

Following Yao (1984), we use

$$c_{ij} = (1-p)^{j-i-1} p, \qquad j < n$$

and

$$= (1-p)^{j-i-1}, \qquad j = n,$$

where $0 \le p \le 1$. This choice implies that the sequence of change points forms a discrete renewal process with interarrival times identically and geometrically distributed.

3.2 Choosing the Block Prior Densities

We propose $f_{ij}(\mu)$ to be the density of $N(\mu_0, \sigma_0^2/(j-i))$. In this way we allow weak signals provided that there are sufficient data available to estimate them. The prior gives higher probability to small departures from μ_0 in large blocks than it does in small blocks; we can expect to identify small departures if they persist for a long time.

Using this model, the density of the observations X_{ij} for a given block ij are

$$f_{ij}(X_{ij}) = \frac{1}{(2\pi\sigma^2)^{(j-i)/2}} \left(\frac{\sigma^2}{\sigma_0^2 + \sigma^2}\right)^{1/2} \exp(V_{ij}), \qquad (1)$$

where

$$V_{ij} = -\frac{\sum_{l=i+1}^{j} (X_l - \bar{X}_{ij})^2}{2\sigma^2} - \frac{(j-i)(\bar{X}_{ij} - \mu_0)^2}{2(\sigma_0^2 + \sigma^2)}$$

with $\bar{X}_{ij} = \sum_{l=i+1}^{j} X_l/(j-i)$.

Given the partition ρ, the estimate of μ_r for $r \in ij \in \rho$ is

$$\hat{\mu}_r = (1 - w)\bar{X}_{ij} + w\mu_0, \tag{2}$$

where $w = \sigma^2/(\sigma_0^2 + \sigma^2)$.

Yao (1984) has considered a similar model, except that $f_{ij}(\mu)$ is the density of $N(\mu_0, \sigma_0^2)$. Given the partition ρ, the estimate of μ_r for $r \in ij \in \rho$ is

$$\tilde{\mu}_r = (1 - w_{ij})\bar{X}_{ij} + w_{ij}\mu_0,$$

where $w_{ij} = \sigma^2/((j - i)\sigma_0^2 + \sigma^2)$.

Both of these models depend on four parameters: p, μ_0, σ_0^2, and σ^2. Only the interpretation of σ_0^2 differs from one case to the other. Yao (1984) proposed choosing these parameters to maximize the likelihood function defined by

$$L(p, \mu_0, \sigma_0^2, \sigma^2) = \sum_{\rho} f[\mathbf{X} \mid \rho] f[\rho],$$

where $f[\mathbf{X} \mid \rho]$ is the probability of the data $\mathbf{X}$ given the partition ρ and $f[\rho]$ is the prior probability of ρ. Because $L(1, \mu_0, \sigma^2, 0) = L(p, \mu_0, 0, \sigma^2)$ for any values of p, μ_0, and σ^2, the model is not identifiable. In response to this Yao proposed maximizing the likelihood subject to the constraint $p \le p_0$, where $p_0 \in (0, 1)$ is a prespecified number. But Yao computed approximations to the posterior means rather than the exact posterior means. In implementing the maximum likelihood approach, we will use the exact computations as outlined in Section 2.

For our model, we pursue a fully Bayesian approach by specifying independent priors for each of the parameters p, μ_0, σ^2, and $w = \sigma^2/(\sigma_0^2 + \sigma^2)$. We let

$$f(\mu_0) = 1, \qquad -\infty \le \mu_0 \le \infty,$$
$$f(\sigma^2) = 1/\sigma^2, \qquad 0 \le \sigma^2 \le \infty,$$
$$f(p) = 1/p_0, \qquad 0 \le p \le p_0,$$

and

$$f(w) = 1/w_0, \qquad 0 \le w \le w_0,$$

where p_0 and w_0 are prespecified numbers in $[0, 1]$. The first two priors are chosen to generate estimates invariant under location and scale changes. The restrictions on the ranges of p and w are designed to make the technique effective in situations where there are not too many changes (p small) and where the changes that do occur are of a reasonable size (w small).

Using this prior, we have that

$$f[\mathbf{X} \mid \rho, \mu_0, w] \propto \int_0^{\infty} \frac{1}{\sigma^2} \prod_{ij \in \rho} f_{ij}(X_{ij})\, d\sigma^2,$$

where $f_{ij}(X_{ij})$ is given by (1). Hence

$$f[\mathbf{X} \mid \rho, \mu_0, w] \propto \frac{w^{b/2}}{[W + Bw + wn(\mu_0 - \bar{X})^2]^{n/2}}, \tag{3}$$

where b = number of blocks in ρ, $\bar{X} = \sum_{i=1}^{n} X_i/n$, $B = \sum_{ij \in \rho}(j - i)(\bar{X}_{ij} - \bar{X})^2$, and $W = \sum_{ij \in \rho} \sum_{l=i+1}^{j} (X_l - \bar{X}_{ij})^2$. Averaging (3) over μ_0 and w gives

$$f[\mathbf{X} \mid \rho] \propto \int_0^{w_0} \frac{w^{(b-1)/2}}{[W + Bw]^{(n-1)/2}}\, dw.$$

Also,

$$f(\rho) = \frac{1}{p_0} \int_0^{p_0} f[\rho \mid p]\, dp$$
$$= \frac{1}{p_0} \left[\int_0^{p_0} p^{b-1}(1 - p)^{n-b}\, dp \right].$$

Hence

$$f[\rho \mid \mathbf{X}] = f[\mathbf{X} \mid \rho] f[\rho]/f[\mathbf{X}]$$
$$\propto \left[\int_0^{w_0} \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}\, dw \right] \left[\int_0^{p_0} p^{b-1}(1 - p)^{n-b}\, dp \right].$$

These integrals are incomplete beta integrals. For $1 \le r \le n$,

$$E[\mu_r \mid \mathbf{X}] = \sum_{\rho} E[\mu_r \mid \mathbf{X}, \rho] f[\rho \mid \mathbf{X}]$$
$$= \sum_{\rho} E[(1 - w)\bar{X}_{ij} + w\mu_0 \mid x, \rho] f[\rho \mid \mathbf{X}].$$

Using (3), it follows after some calculation that

$$E[\mu_0 \mid w, \mathbf{X}, \rho] = \bar{X}$$

and

$$E[w \mid \mathbf{X}, \rho] = w^*,$$

where

$$w^* = \frac{\int_0^{w_0} \frac{w^{(b+1)/2}}{[W + Bw]^{(n-1)/2}}\, dw}{\int_0^{w_0} \frac{w^{(b-1)/2}}{[W + Bw]^{(n-1)/2}}\, dw}.$$

Hence

$$E[\mu_r \mid \mathbf{X}] = \sum_{\rho} [(1 - w^*)\bar{X}_{ij} + w^*\bar{X}] f[\rho \mid \mathbf{X}].$$

Likewise,

$$E[\sigma^2 \mid \mathbf{X}] = \sum_{\rho} E[\sigma^2 \mid \rho, \mathbf{X}] f[\rho \mid \mathbf{X}].$$

Using the fact that

$$f[\sigma^2, \mu_0, w \mid \rho, \mathbf{X}] \propto \frac{1}{\sigma^2} \prod_{ij \in \rho} f_{ij}(X_{ij}),$$

we get after integration that

$$E[\sigma^2 \mid \rho, \mathbf{X}] = \frac{1}{n - 3} \frac{\int_0^{w_0} \frac{w^{(b-1)/2}}{[W + Bw]^{(n-3)/2}}\, dw}{\int_0^{w_0} \frac{w^{(b-1)/2}}{[W + Bw]^{(n-1)/2}}\, dw}.$$

It should be noted that $f[\rho \mid \mathbf{X}]$ is not in product form, and so the iterative calculations described in Section 2 are not possible here. In the next section we describe how Markov sampling may be used to achieve a satisfactory approximation to $E[\mu_r \mid \mathbf{X}]$.

Chernoff and Zacks (1964) considered a slightly different model in the context of estimating μ_n, the "current mean," given data $X_1, X_2, \ldots, X_n$ with X_i distributed as $N(\mu_i, \sigma^2)$.

They assumed that the sequence of parameter values $\mu_1, \mu_2, \ldots, \mu_n$ forms a Markov chain with $\mu_{i+1} = \mu_i + \delta_i$ where $\delta_i = 0$ with probability $1 - p$ and δ_i is distributed as $N(0, \sigma_0^2)$ with probability p. Hence, writing $\delta_1 = 0$, we have

$$\mu_i = \mu_1 + \sum_{j=1}^{i} \delta_j$$

for $i = 1, 2, \ldots, n$. For convenience we parameterize in terms of $\lambda = \sigma^2/\sigma_0^2$ and σ^2 rather than σ_0^2 and σ^2.

Writing $\delta = (\delta_1, \delta_2, \ldots, \delta_n)$, we have for the partition $\rho = (i_0, i_1, \ldots, i_b)$,

$$f(\rho, \delta, \mu_1, \sigma^2, \lambda, p \mid \mathbf{X}) \propto \frac{p^{k-1}(1-p)^{n-k}\lambda^{(k-1)/2}}{\sigma^{n+k-1}} \times \exp\left(-\frac{G}{2\sigma^2}\right) f(\mu_1) f(\lambda) f(\sigma^2) f(p),$$

where $f(\mu_1), f(\lambda), f(\sigma^2)$, and $f(p)$ are the prior densities and

$$G = \sum_{i=1}^{n} \left(X_i - \mu_i - \sum_{j=1}^{i} \delta_j\right)^2 + \lambda \sum_{i=1}^{n} \delta_i^2.$$

Following Chernoff and Zacks (1964), we set $f(\mu_1) = 1, -\infty \le \mu_1 \le \infty$. As before, we set $f(\sigma^2) = 1/\sigma^2, 0 \le \sigma^2 \le \infty$ and $f(p) = 1/p_0, 0 \le p \le p_0$, where p_0 is a constant. For λ we consider the prior $f(\lambda) = A \exp(-A\lambda), 0 \le \lambda \le \infty$, where A is a constant. Using these priors, we can integrate over p and μ_1 to get

$$f(\rho, \delta, \sigma^2, \lambda \mid \mathbf{X}) \propto \left[\int_0^{p_0} p^{b-1}(1-p)^{n-b}\, dp\right] \frac{\lambda^{(b-1)/2}}{\sigma^{n+b}} \exp(-G_1/2\sigma^2), \quad (4)$$

where

$$G_1 = \sum_{i=1}^{n} \left(X_i - \sum_{j=1}^{i} \delta_j\right)^2 - \left(\sum_{i=1}^{n} \left(X_i - \sum_{j=1}^{i} \delta_j\right)\right)^2 \Big/ n + \left(\sum_{i=1}^{n} \delta_j^2 + A\right)\lambda. \quad (5)$$

Given $\rho = (i_0, i_1, \ldots, i_b)$ and λ, the posterior means $\hat{\mu}_{i_1}, \hat{\mu}_{i_2}, \ldots, \hat{\mu}_{i_b}$ can be calculated by iteratively solving the system of equations

$$(i_1 + \lambda)\hat{\mu}_1 - \lambda\hat{\mu}_2 = \sum_{l=1}^{i_1} X_l,$$

$$-\lambda\hat{\mu}_{r-1} + (i_r - i_{r-1} + 2\lambda)\hat{\mu}_r - \lambda\hat{\mu}_{r+1} = \sum_{l=i_{r-1}+1}^{i_r} X_l, \qquad r = 2, 3, \ldots, b-1,$$

and

$$-\lambda\hat{\mu}_{b-1} + (n - i_{b-1} + \lambda)\hat{\mu}_b = \sum_{l=i_{b-1}+1}^{n} X_l. \quad (6)$$

In the next section we demonstrate how Markov sampling may be used to calculate $E(\mu_i \mid \mathbf{X})$ by repeated sampling from (3).

4. MARKOV SAMPLING

4.1 The General Technique

The Markov sampling technique originated in statistical physics (Hammersley and Handscomb 1964; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) and more recently has found wide application in image processing (Cross and Jain 1983; Geman and Geman 1984). To sample from a finite random variable Z having density f, we devise a Markov chain with transition probabilities $f(Y \mid Z)$ chosen so that the corresponding stationary probabilities are $f(Z)$. We begin with some arbitrary starting value Z_0 and compute the Markov sequence $Z_1, Z_2, \ldots, Z_m$ according to the transition probabilities $f(Y \mid Z)$. Under certain conditions, the frequency of occurrence of the states Z_i converges to the stationary distribution $f(Z)$.

We will use a transition matrix based on conditional distributions that appeared in Geman and Geman (1984). (They sampled from Gibbs distributions and called their method the Gibbs sampler, but it seems unfortunate to give that name to a method that can be used for sampling from any kind of distribution.)

To explain the technique, assume that U and V are finite variables with a given joint density $f(u, v)$ and conditional densities $f(u \mid v)$ and $f(v \mid u)$. We begin with some initial states U^0, V^0; a transition matrix for generating new states U^i, V^i from states U^{i-1}, V^{i-1} is defined as follows:

$$f(u^i, v^i \mid u^{i-1}, v^{i-1}) = f(u^i \mid v^{i-1}) f(v^i \mid u^i).$$

Thus we generate U^i from the conditional density of U given $V = v^{i-1}$ and then generate V^i from the conditional density of V given $U = u^i$. The joint density $f(u, v)$ is a stationary distribution for this transition matrix, as can be seen by algebra or by supposing that the initial states have the density $f(u, v)$; then V^0 has the correct marginal density and U^1 has the correct conditional density given V^0, so U^1, V^0 has the correct joint density. Continuing one more step, U^1, V^1 has the correct joint density, which is the same density as U^0, V^0, so $f(u, v)$ is indeed a stationary distribution. For a finite chain, the stationary distribution is unique if the chain is *irreducible*; that is, if there exists a path of transitions of positive probability between any two states of the chain. For our problem, we require that for each pair (u, v) and (u', v') of positive probability there exists a sequence $u^1 = u, v^1 = v, u^2, v^2, \ldots, u^n = u'$, and $v^n = v'$ such that $f(u^i, v^{i-1}) > 0$ and $f(u^i, v^i) > 0$.

When the stationary distribution is unique, the limiting distribution of the sampled points U^i, V^i is just the required density $f(u, v)$. More generally, if we have n variables $U_1, \ldots, U_n$ such that each state of positive probability can be reached from each other such state by transitions of positive probability, then we generate a sample with limiting distribution $f(u_1, u_2, \ldots, u_n)$ by beginning with some $u_1^0, \ldots, u_n^0$ and making the transitions one variable at a time conditional on the values of the other variables. We select U_k^i from the density $f(u_k \mid u_1^i, \ldots, u_{k-1}^i, u_{k+1}^{i-1}, \ldots, u_n^{i-1})$.

How does this apply to generating a distribution over partitions? For n observations, let $U_i = 1$ if $\theta_i = \theta_{i+1}$ and let $U_i = 0$ if $\theta_i \ne \theta_{i+1}$, for $1 \le i < n$. Then the transition at U_i is

to 1 or 0, according to the probabilities of partitions $f(U_1, \ldots, U_{i-1}, 1, U_{i+1}, \ldots, U_{n-1})$ and $f(U_1, \ldots, U_{i-1}, 0, U_{i+1}, \ldots, U_{n-1})$.

4.2 Calculation of the Bayes Estimate

A pass through the data selects for each i in turn a value for U_i from the distribution of U_i given $\mathbf{X}$ and the current values of $U_j, j \neq i$. We initialize all U_i to be 0.

Consider position $i \geq 2$. Let b denote the number of blocks we obtain if we set $U_i = 0$. Then

$$\frac{P(U_i = 1 \mid \mathbf{X}, U_j, j \neq i)}{P(U_i = 0 \mid \mathbf{X}, U_j, j \neq i)} = \frac{[\int_0^{p_0} p^b(1-p)^{n-b-1}\,dp]\left[\int_0^{w_0} \frac{w^{b/2}}{(W_1 + B_1 w)^{(n-1)/2}}\,dw\right]}{[\int_0^{p_0} p^{b-1}(1-p)^{n-b}\,dp]\left[\int_0^{w_0} \frac{w^{(b-1)/2}}{(W_0 + B_0 w)^{(n-1)/2}}\,dw\right]},$$

where W_0 is the within-block sum of squares and B_0 is the between-block sum of squares we obtain with $U_i = 0$; W_1 and B_1 are similarly defined. Hence we can calculate $f(U_i = 1 \mid \mathbf{X}, U_j, j \neq i)$ and so sample U_i at random.

Having completed a pass through the data, we now calculate the posterior mean of μ_i given the current partition ρ and $\mathbf{X}$ using (2). At the end of M passes, we use the average of the M estimates of μ_i as an approximation to the posterior mean of μ_i given $\mathbf{X}$. Likewise, we calculate the posterior mean of σ^2 given ρ and $\mathbf{X}$ and use the average over the M passes as an approximation to the posterior mean of σ^2 given $\mathbf{X}$. We find that $M = 50$ to 500 gives adequate approximations to the posterior mean.

4.3 Calculation of the Chernoff–Zacks Estimate

We wish to sample from $f(\delta, \sigma^2, \lambda \mid \mathbf{X})$ given by (4). We initially set $\lambda = 1$ and $\delta_i = 0$, $i = 1, 2, \ldots, n$.

Each pass through the data consists of generating a value for σ^2 given λ and δ, then a value for λ given σ^2 and δ, and then for each $i \geq 2$ a value for δ_i given σ^2, λ, and $\delta_j, j \neq i$. Given λ and δ, $G_1/\sigma^2 \sim \chi^2_{n+b-2}$, where b is the number of blocks in the partition and G_1 is given by (5). Given σ^2 and δ, $(\sum_{i=2}^n \delta_i^2/\sigma^2 + 2A)\lambda \sim \chi^2_{b+1}$. Given σ^2, λ, and $\delta_j, j \neq i$, let b be the number of blocks obtained if we set $\delta_i = 0$. Then $\delta_i = 0$ with probability $1 - p_i$, and δ_i is distributed as $N(\alpha_i, \beta_i^2)$ with probability p_i, where

$$\alpha_i = \frac{(i-1)\sum_{k=i}^n X_k^* - (n-i+1)\sum_{k=1}^{i-1} X_k^*}{n\lambda + (i-1)(n-i+1)},$$

$$\beta_i = \frac{n\sigma^2}{n\lambda + (i-1)(n-i+1)},$$

$$\frac{p_i}{1-p_i} = \frac{\int_0^{p_0} p^b(1-p)^{n-b-1}\,dp}{\int_0^{p_0} p^{b-1}(1-p)^{n-b}\,dp} \times \left[\frac{n\lambda}{n\lambda + (i-1)(n-i+1)}\right]^{1/2} \exp(\alpha_i^2/2\beta_i^2),$$

and

$$X_k^* = X_k - \sum_{j=1}^{k} \delta_j, \qquad k < i$$

$$= X_k - \sum_{j=1}^{i-1} \delta_j - \sum_{j=i+1}^{k} \delta_j, \qquad k \geq i.$$

After each pass, the posterior means of μ_i given the current ρ and λ are computed using (6). At the end of M passes, we use the average of the M estimates of μ_i as an approximation to the posterior mean given the observations. Likewise, we use the average of the M-generated values of σ^2 as an approximation to the posterior mean of σ^2 given $\mathbf{X}$. The general approach is similar to the EM algorithm, in that we identify conditioning variables ρ and λ for which the conditional posterior means are easy to compute, and then average these conditional means over sampled values of ρ and λ to get an estimate of the posterior means conditioned only on the observations.

5. THE DESIGN OF THE SIMULATION STUDY

A Monte Carlo study was carried out to assess the effectiveness of the Barry–Hartigan estimate and to compare it with the Yao method, the Chernoff–Zacks method, and a method based on the Schwarz criterion. Evaluations were carried out for various partitions and parameter values μ; each partition and parameter sequence is called a scene.

5.1 The Estimators

Four methods of estimation were considered:

1. B–H: the Bayesian estimate of Section 3 with $p_0 = .2$ and $w_0 = .2$. These choices of (p_0, w_0) were made after extensive experimentation over various scenes.

2. YAO: Yao's method with $p_0 = .2$, the value proposed in Yao (1984). We experimented with other values for p_0 but found that $p_0 = .2$ performed best over a wide variety of scenes.

3. C–Z: The method of Chernoff and Zacks with $p_0 = .2$ and $A = 1.0$; the value $A = 1$ means that the expected value of the ratio of the error variance σ^2 to the parameter jump variance σ_0^2 is 1. Other values of A did not change the behavior of the method very much.

4. SCHWARZ: A method based on Schwarz's criterion described in Yao (1988), which chooses a partition ρ to maximize $-n \log \hat{\sigma}_\rho - b \log n$, where b is the number of blocks in ρ and $\hat{\sigma}_\rho$ is the maximum likelihood estimate of σ given ρ. Having selected ρ, the μ's are estimated using the observed block means. In using this method we restricted the search for a maximum to those partitions with fewer than $n/2$ blocks.

5.2 The Scenes

Given values for n, p, μ_0, w, and σ^2, we generate scenes as follows. For $1 < i \leq n$, the probability that a change occurs at position i is p independently of what occurs at other sites. Given the blocks of the partition ρ so generated, the mean value μ_j for block ij is chosen at random from the density of $N(\mu_0, \sigma_0^2/(j-i))$, where $\sigma_0^2 = \sigma^2(1-w)/w$. The data

Table 1. Comparison of Methods for 20 Scenes

	Scene	B–H	YAO	C–Z	SCHWARZ
One Group					
1.	60°	2.40^{29}	1.84^{12}	1.75^{18}	3.05^{41}
Two Groups					
2.	$40^0\ 20^3$	2.24^{18}	2.17^{8}	3.74^{13}	3.04^{27}
3.	$40^0\ 20^2$	2.42^{17}	2.37^{8}	2.76^{11}	2.77^{23}
4.	$30^0\ 30^1$	2.63^{18}	2.82^{8}	2.26^{7}	3.58^{22}
5.	$30^0\ 30^{0.5}$	2.04^{14}	2.26^{7}	1.78^{6}	3.16^{16}
6.	$58^0\ 2^3$	2.90^{21}	5.16^{29}	3.17^{11}	3.19^{28}
Three Groups					
7.	$15^0\ 30^2\ 15^0$	2.77^{14}	2.91^{5}	2.62^{8}	3.84^{19}
8.	$10^0\ 40^1\ 10^0$	2.60^{10}	2.70^{6}	2.39^{4}	4.40^{13}
9.	$30^0\ 20^2\ 10^0$	3.08^{18}	3.13^{6}	3.19^{9}	3.40^{20}
10.	$4^0\ 1^5\ 55^0$	3.32^{22}	6.86^{29}	7.61^{21}	3.09^{18}
Four Groups					
11.	$10^1\ 15^0\ 20^2\ 15^1$	2.46^{9}	2.40^{4}	2.39^{6}	3.69^{12}
12.	$10^1\ 10^2\ 10^1\ 30^0$	2.21^{9}	2.34^{3}	1.97^{4}	3.54^{11}
13.	$25^2\ 1^0\ 14^5\ 20^3$	2.46^{12}	2.68^{4}	3.81^{8}	2.83^{11}
14.	$15^3\ 5^0\ 5^5\ 35^3$	2.48^{15}	3.02^{4}	6.89^{27}	2.74^{16}
Five Groups					
15.	$5^0\ 5^2\ 40^0\ 5^2\ 5^0$	3.17^{12}	3.29^{9}	4.03^{8}	4.97^{14}
16.	$12^0\ 12^1\ 12^0\ 12^1\ 12^0$	2.21^{7}	2.10^{4}	2.18^{4}	3.62^{7}
17.	$12^0\ 12^1\ 12^2\ 12^3\ 12^4$	2.29^{7}	2.28^{3}	1.85^{4}	3.87^{10}
Eight Groups					
18.	$14^3\ 5^5\ 1^0\ 1^2\ 4^5$ $9^2\ 15^3\ 11^4$	2.48^{12}	2.61^{5}	5.23^{10}	3.87^{14}
19.	$2^1\ 19^0\ 20^1\ 15^4\ 5^1$ $1^5\ 5^0\ 3^1$	2.40^{9}	2.74^{6}	4.37^{9}	3.32^{11}
Ten Groups					
20.	$6^0\ 6^2\ 6^0\ 6^2\ 6^0$ $6^2\ 6^0\ 6^2\ 6^0\ 6^2$	2.99^{9}	2.85^{8}	4.88^{8}	5.26^{11}

NOTE: For each method we give the mean sum of squares per block over 100 repetitions. The superscript is the standard error of the mean for B–H; for the other methods it is the standard erorr of the difference between the mean and the mean for B–H.

values X_{ij} are then generated by adding independent $N(0, \sigma^2)$ errors to μ_j.

Throughout we use $n = 60$, $\mu_0 = .0$, and $\sigma^2 = 1.0$. A number of scenes were generated by choosing p's at random from $U(0, .1)$ and w's from $U(0, .2)$. The scenes chosen for inclusion in the study, shown in Table 1, are representative of those generated. We use the notation $5^0\ 5^2\ 40^0\ 5^2\ 5^0$, for example, to denote a partition into five blocks of lengths 5, 5, 40, 5, and 5 and parameter values 0, 2, 0, 2, and 0. Scenes numbered 16 and 20 were included to favor Yao; scene number 17, to favor C–Z. For scenes chosen according to a certain prior distribution, the posterior Bayes means according to that prior must have minimum mean squared error (MSE) compared to any estimate, when the errors are averaged over that prior distribution. We thus expect the B–H method to be tautologically superior to the others when averaged over the scenes sampled from the prior. It is nevertheless of interest to compare the MSE's of the different methods in each scene, to identify situations where the various methods are superior.

The values of $p_0 = .1$ and $w_0 = .2$ were used to generate scenes with relatively few blocks and relatively large jumps. The inconsistency between the values used in generating the scenes and those used in the B–H method arises because we discovered a need to overfit the number of blocks to produce low MSE's in the posterior means. If there is a relatively small difference between a block mean and its neighboring means, then it is likely that the block will not be detected and might make a substantial contribution to the MSE's. To prevent this, we increase the likelihood of detection of slightly different blocks, while at the same time increasing the chance of discovering false blocks and so overestimating the number of blocks. Therefore, we encourage more blocks by allowing p and w to take larger values in the fitting than in the construction. In practice we would constrain, say $p < .2$, even though we do not anticipate blocks of average length 5; this permits somewhat more blocks in the fitting than we really believe are there.

5.3 Implementation

For each of the 20 scenes chosen, 100 sets of data were generated. For each of the four estimates, we calculated the sum of squared errors per block (SSPB), defined as the sum of squared errors divided by the number of blocks in the true partition. If the true partition were known with certainty, the estimate within each block would be the block mean, the contribution of each block to the expected sum of squared errors would be 1, and the total expected sum of squared errors would be the number of blocks. Thus the SSPB is a standardized measure of error for each scene, which would ideally be 1 if we could surely identify the correct partition.

For the Bayesian methods, we found that $M = 110$ passes through the data, omitting the first 10 passes from the estimation process, gave accurate approximations to the posterior means. For the C–Z method, we found that more passes were required for accurate determination of posterior means. This is largely due to a greater degree of autocorrelation between passes, caused by the need to sample δ at each pass. After some experimentation, we used $M = 510$ and omitted the first 10 passes from the estimation process.

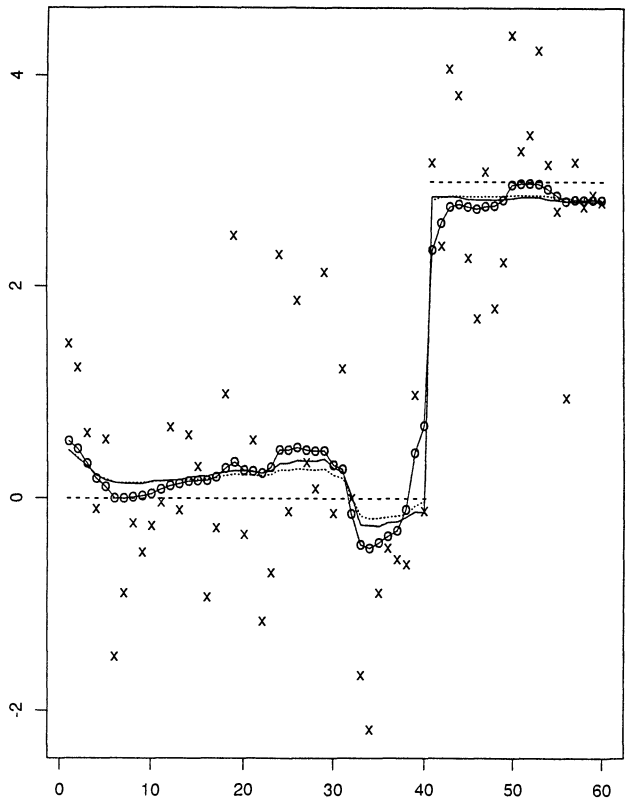

Figure 1. Fitted Values in a Particular Sample for Scene 2. - - -, True; ×, Data; ——, B–H; -----, YAO; -O-, C–Z.

5.4 Bias Adjustments for Markov Sampling

The estimates based on Markov sampling converge to the posterior means as the number of passes, M, increases. How should M be chosen? In the following analysis, the observations are regarded as fixed; we consider only error due to the Markov sampling. Let $\hat{F}_i = E(\mu_i|x)$ be the posterior mean of μ_i and let $\hat{F}_{Mi}$ be the estimate of $\hat{F}_i$ based on M passes through the data. Then, for M large,

$$\hat{F}_{Mi} = \hat{F}_i + \eta_i,$$

where η_i has mean $O(1/M)$ and variance proportional to $1/M$. The mean is $O(1/M)$, because $\hat{F}_{Mi}$ is the average of M conditional means given the M sampled partitions, and these conditional means have average value that approaches $\hat{F}_i$ exponentially fast as $M \to \infty$. This follows because the k step transition matrix of a Markov chain converges exponentially fast to the stationary distribution in all rows. (We also reduce the bias in η_i by ignoring the first few Markov samples.) The assumption of variance proportional to $1/M$ derives from the approximate independence of the sampled ρ in well-separated Markov passes. Hence writing R for the SSPB using $\hat{F}_i$ and R_M for the SSPB using $\hat{F}_{Mi}$, we have

$$E(R_M) = E(R) + B/M + O(M^{-2}),$$

where B is a constant depending on n. To correct for this bias, we recorded for the kth data set the SSPB $R^1_{k,M/2}$ for estimates based on the first $M/2$ passes, the SSPB $R^2_{k,M/2}$ for estimates based on the second $M/2$ passes, and the SSPB $R_{k,m}$ for estimates based on all M passes. The jackknifed quantity

$$\hat{R}_{adj} = \sum_{k=1}^{100} [2R_{k,M} - (R^1_{k,M/2} + R^2_{k,M/2})/2]/100$$

is a less-biased estimate of $E(R)$, the expected value of SSPB if the exact posterior means had been used. To get an idea of the magnitude of the bias, we also recorded the unadjusted estimate of $E(R)$ given by $\hat{R} = \sum_{k=1}^{100} R_{k,M}/100$. We took M large enough so that the differences between the adjusted and unadjusted estimates of MSE were negligible.

There is also an increase in the variance of R due to the Markov sampling. For M large, this increase is of order $1/M$, and we estimated it by $[\sum_{k=1}^{100} (R^1_{k,M/2} - R^2_{k,M/2})^2]/(2)(2)(100)$.

6. THE RESULTS OF THE SIMULATION STUDY

The results of the simulation are displayed in Table 1. For YAO and SCHWARZ we report the mean SSPB; for B–H and C–Z we report the bias-adjusted mean SSPB. The standard error (SE), given as a superscript, is the SE of the mean for B–H and for each of the other methods is the SE of the difference between the mean and the mean for B–H. For ease of presentation, we use a superscript of 29 to denote a standard error of .29.

Computations on a 20 mHz PC for a single sample from scene 15 took 121 seconds for B–H, 342 seconds for YAO, 162 seconds for C–Z, and 5 seconds for SCHWARZ. It should be noted that Yao (1984) recommended an approximate technique for estimating posterior means that would considerably improve the given time.

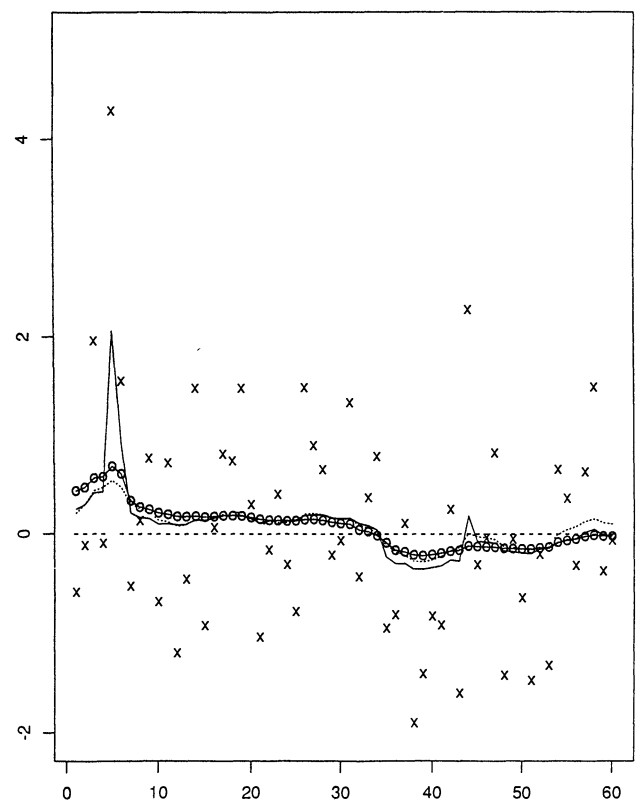

Figure 2. Fitted Values in a Particular Sample for Scene 10. - - -, True; ×, Data; ——, B–H; -----, YAO; -O-, C–Z.

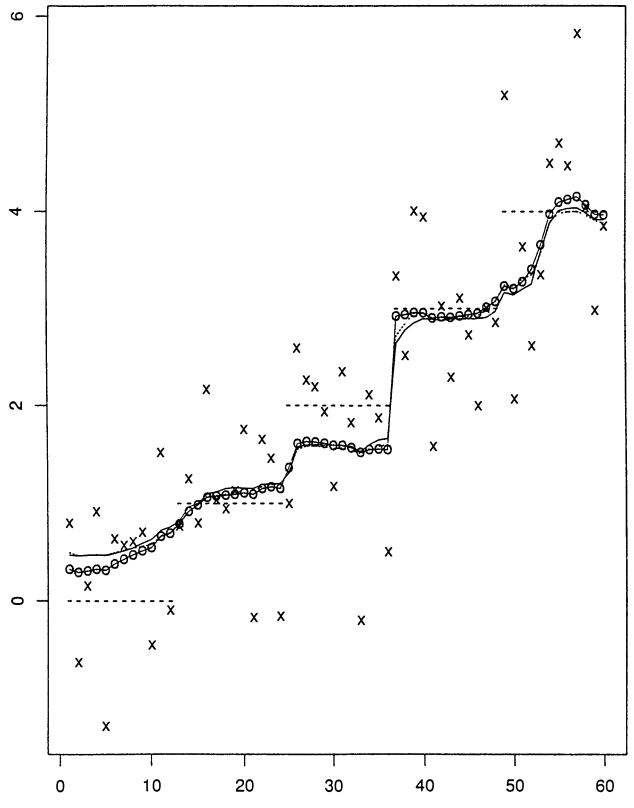

Figure 3. Fitted Values in a Particular Sample for Scene 17. – – –, True; ×, Data; ——, B–H; -----, YAO; -O-, C–Z.

The bias corrections for both B–H and C–Z were quite small, ranging from 0 to .10 for B–H and from 0 to .15 for C–Z. In addition, the increase in the SE of the mean SSPB caused by the Markov sampling was small—less than 10% in all cases and usually much less than that. This indicates that for each method, the number of passes used was sufficient to allow accurate comparison with other methods. In a practical application of either method, it would be sensible to use more passes in the Markov sampling than was feasible in a large-scale simulation study such as this; for example, use $M = 500$ rather than $M = 100$.

The most obvious conclusion from Table 1 is that SCHWARZ compares quite poorly with the three Bayesian methods with specific priors, but does better than YAO or C–Z for short sharp signals. Because SCHWARZ method is forced to opt for one partition, it either does very well when it selects correctly or very badly when it selects incorrectly. In contrast, the Bayesian methods allow for uncertainty about which is the correct partition by weighting various partitions according to how much evidence in their favor the data contain.

In Figures 1–4 we show plots of the fitted values for examples from Scenes 2, 10, 17, and 20. Each plot shows the underlying scene, the raw data, and the fitted values for B–H, YAO, and C–Z.

C–Z works well when the changes are small and persistent but poorly when changes are large or persist for only a few observations. YAO works well when changes occur regularly and when the sizes of the changes do not exhibit much variation, but does not handle outliers well. B–H is nearly always close in SSPB to the other techniques and does considerably better when the underlying scene consists of irregularly distributed changes of variable magnitudes; it also handles outliers better than either of the other methods.

The behavior of C–Z is well illustrated by the figures. In Figure 1 (scene 2) we see that C–Z is slower to react to sharp changes than either of the other methods. In Figure 4 (scene 20) we see how C–Z fails to pick up changes that last for only a few observations; however, the results for scenes 4 and 5 indicate that C–Z reacts better than the other two Bayesian methods to small changes that persist for a long time. It does slightly better when the scene consists of gradual changes, such as scene 17; however, Figure 3 (scene 17) illustrates the similarity of the three estimates in this case.

The difference between YAO and B–H is well illustrated by a comparison of Figure 2 (scene 10) and figure 4 (scene 20). YAO does well for scene 20, where the true scene consists of 10 blocks all of length 6 and with all jumps being equal to 2; however, it fails completely to react to the short sharp change of scene 10. Because the block prior densities for YAO are identical across all blocks, YAO cannot cope with scenes that suggest that different prior densities are appropriate in different blocks. In scene 10 YAO would need to use a large value of σ_0^2 to react appropriately to the short sharp jump and a small value of σ_0^2 to do sufficient smoothing elsewhere; it opts for a small value and moves all estimates towards the overall mean. B–H, on the other hand, automatically scales the prior variance because it uses $\sigma_0^2/(j - i)$ for block ij, which allows it to use simultaneously a large variance for short blocks and a small variance for long blocks.

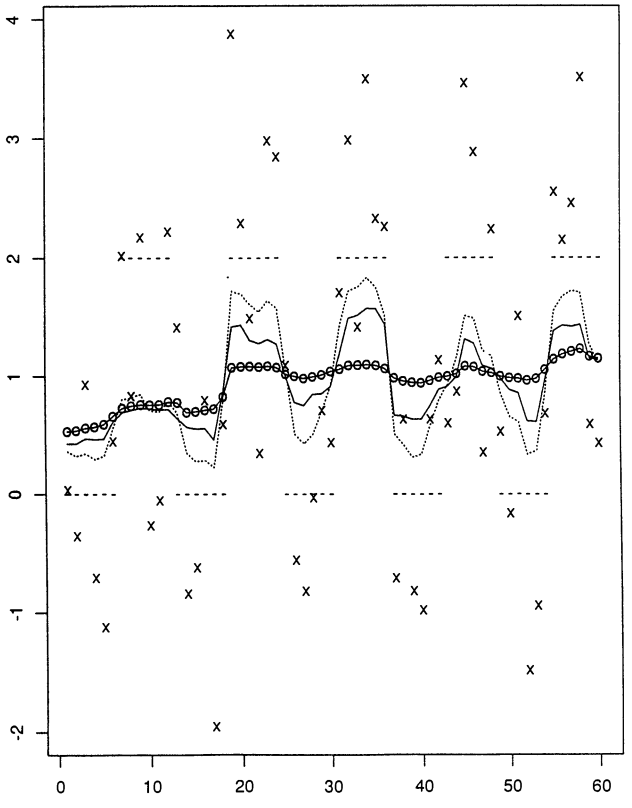

Figure 4. Fitted Values in a Particular Sample for Scene 20. – – –, True; ×, Data; ——, B–H; -----, YAO; -O-, C–Z.

B–H can, therefore, be sensitive to short sharp changes without having to be overly sensitive to mere random deviations in long blocks.

There is, however, a price to pay for the extra sensitivity of B–H. This can be seen most clearly in the results for scene 1, the case of no changes. In 84 of the 100 repetitions YAO fit the sample mean, whereas B–H fit the mean in only 12 of the repetitions. C–Z fit the mean for only 19 of the repetitions but rarely reacted very strongly to random deviations in the data and so performed well in MSE terms. B–H was overly sensitive to random fluctuations in the data, mistaking them for changes.

Looking at the results overall, we conclude that B–H is preferable in a minimax sense, because it never produces the disastrously large MSE's of the other methods and pays only a modest price for this degree of safety.

7. APPLICATION

We illustrate the techniques compared in the simulation study using data of Lombard (1987), which give the radii of 100 successive circular indentations cut by a milling machine. Figure 5 shows the raw data and the estimated means produced by the three Bayesian techniques. SCHWARZ fits one mean, yielding a value of .01005 as an estimate of the process variability σ^2. The three Bayesian techniques produce quite similar results. The phenomena described by Lombard—an increase between 20 and 40 and of a decrease at observation 76—are evident. In addition, there is evidence of a dip in mean value around observation 55 and a steady increase after observation 84. The estimates of σ^2 were .00857 for B–H, .00835 for YAO, and .00898 for C–Z.

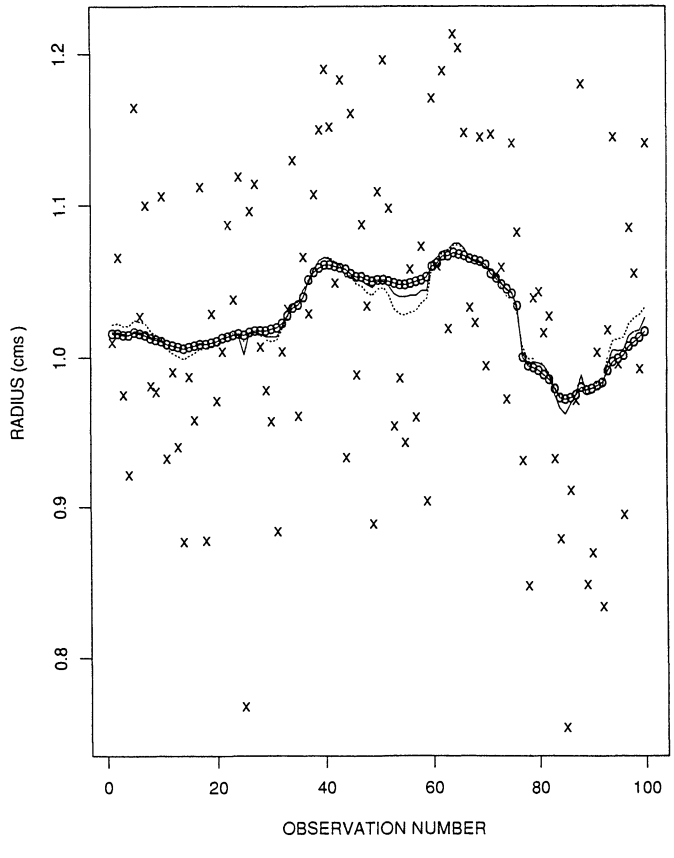

Figure 5. Comparison of Fits From Bayesian Methods Applied to Lombard's Data. ×, Data; ——, B–H; -----, YAO; –○–, C–Z.

Even though the change points are not distinct in this application, the Bayesian models express appropriate uncertainty about the change points and give a plausible adaptive smoothing of the original observations. The estimate from C–Z is smoother than that from either YAO or B–H. The B–H estimate agrees closely with the C–Z estimate except at observations 25, 85, and 88. These might be considered outlying observations; the B–H estimate, being sensitive to short sharp changes in the underlying means, reacts to them as if a change in mean might have occurred. The reaction is localized in the sense that only the fitted value corresponding to the suspected outlier is affected. The YAO estimate reacts less sharply to these suspected outliers, at least in a local sense, but is generally more variable than the B–H estimate. This is due to the fact that the maximum likelihood choice for the parameter p in this case was $p = .15$. Outliers encourage high values of p and so have a global effect on the Yao estimate, in contrast to their sharp local effect on the B–H estimate.

8. EXTENSIONS

The Bayesian methodology developed in this article can be extended to provide solutions to many other changepoint-like problems. Allowing positive prior cohesions for connected subsets only is appropriate for time-ordered observations, for example. But there is no intrinsic difficulty in extending the methodology to allow positive cohesions for all subsets. There would be an increase in the computational burden, but this increase would be somewhat mitigated through Markov sampling. The methodology also can be extended to observations made in two or higher dimensions and thus offers solutions to many image processing problems. In image processing we might consider allowing positive prior cohesions for connected subsets only, for convex subsets only, or indeed for any family of subsets with the cohesions chosen to encourage desirable properties such as connectivity. We intend to explore this idea further in a forthcoming article.

In an earlier article (Barry and Hartigan 1990), we described how product partition models can be applied to problems involving changing regression regions. The ideas of this article can easily be extended to such problems. The methodology can also be applied when the data are not normally distributed and/or are not independent given the parameter values. The computational difficulty will depend on the complexity of the posterior distributions that arise. When applying Markov sampling, it is good to integrate the posterior probabilities of partitions over as many of the unknown parameters as possible before sampling partitions. This reduces the correlation between the simulated posterior means given the sampled partition in successive passes, so that fewer passes are required to get satisfactory approximations for the posterior means. With the aid of Markov sampling, we expect product partition models to be useful in problems in which there is uncertainty concerning the range of application of a particular probability model.

[*Received October 1990. Revised March 1992.*]

REFERENCES

Barry, D., and Hartigan, J. A. (1992), "Product Partition Models for Change Point Problems," *The Annals of Statistics,* 20, 260–279.

Chernoff, H., and Zacks, S. (1964), "Estimating the Current Mean of a Normal Distribution Which is Subject to Changes in Time," *Annals of Mathematical Statistics,* 35, 999–1018.

Cross, G. C., and Jain, A. K. (1983), "Markov Random Field Texture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-5, 25–39.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-6, 721–741.

Hammersley, J. M., and Handscomb, D. C. (1964), *Monte Carlo Methods,* London: Methuen.

Hartigan, J. A. (1990), "Partition Models," *Communications in Statistics,* 19, 2745–2756.

Lombard, F. (1987), "Rank Tests for Changepoint Problems," *Biometrika,* 74, 615–624.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics,* 21, 1087–1091.

Yao, Y. C. (1984), "Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches," *Annals of Statistics,* 12, 1434–1447.

——— (1988), "Estimating the Number of Change-Points by Schwarz's Criterion," *Statistics and Probability Letters,* 6, 181–187.