The Annals of Statistics
1992, Vol. 20, No. 1, 260–279

PRODUCT PARTITION MODELS FOR CHANGE POINT PROBLEMS[1]

BY DANIEL BARRY AND J. A. HARTIGAN

University College, Cork, and Yale University

Product partition models assume that observations in different components of a random partition of the data are independent. If the probability distribution of random partitions is in a certain product form prior to making the observations, it is also in product form given the observations. The product model thus provides a convenient machinery for allowing the data to weight the partitions likely to hold; and inference about particular future observations may then be made by first conditioning on the partition and then averaging over all partitions. These models apply with special computational simplicity to change point problems, where the partitions divide the sequence of observations into components within which different regimes hold. We show, with appropriate selection of prior product models, that the observations can eventually determine approximately the true partition.

1. Introduction. In one-dimensional change point problems, the sequence of observations $X_1, \ldots, X_n$ observed at consecutive points in time, is partitioned into b contiguous subsequences or *blocks*,

$$\left[X_1, \ldots, X_{i_1}\right], \left[X_{i_1+1}, \ldots, X_{i_2}\right], \ldots, \left[X_{i_{b-1}+1}, \ldots, X_{i_b}\right],$$

and a different probability model is assumed to hold within each of the blocks. Part of the inference problem is specifying the partition into blocks.

We will assume that the partition is randomly selected according to a *product partition* distribution: The probability of the partition ρ into blocks ending at $i_1, i_2, \ldots, i_b$ is

$$p(\rho) = Kc_{0i_1}c_{i_1i_2} \cdots c_{i_{b-1}i_b}$$

for some assignment of nonnegative cohesions c_{ij} to the block of observations between $i + 1$ and j for each $0 \le i < j \le n$.

Equivalently, the endpoints $0 = i_0, i_1, i_2, \ldots, i_n = n$ of the blocks form a Markov chain satisfying

$$0 \le i_r < i_{r+1} \le n \quad \text{for } i_r < n,$$
$$i_r = i_{r+1} = n \quad \text{for } i_r = n.$$

Product partition distributions are defined for general random partitions, not just ones consisting of contiguous subsequences, in Hartigan (1990). They are called product distributions because the probability of a partition ρ into

Received May 1989; revised May 1991.
[1]Research partially supported by NSF Grant DMS-86-17919.
AMS 1980 *subject classifications*. Primary 62M20, 93E14; secondary 62G05, 62C12.
Key words and phrases. Product partition models, change points.

subsets $S_1, S_2, \ldots, S_b$ may be written as the product

$$p(\rho) = Kc(S_1)c(S_2) \cdots c(S_b)$$

for some set of nonnegative *cohesions* $c(S)$ defined for each subset S of the points $1, 2, \ldots, n$. The constant K is chosen so that the sum over all partitions is unity.

In the present case, the cohesions are zero for all sets not consisting of contiguous points in time. Let ij denote the set of points $i + 1, \ldots, j$ for $i < j$ and let p_{ij} denote the probability that the Markov chain of endpoints makes a transition to j given i. Allow $p_{nn} = 1$, to take care of terminating the chain. A partition ρ is identified by the set of block endpoints $i_0 = 0, i_1, i_2, \ldots, i_n$; the number of blocks $b = b(\rho)$ is obtained by finding b such that $i_{b-1} < n$, $i_b = n$. The set of transition probabilities is a set of cohesions for the product partition distributions on contiguous sets:

$$p(\rho) = p_{0i_1}p_{i_1i_2} \cdots p_{i_{b-1}i_b}.$$

An important quantity in practical computations is the *relevance* $r(ij)$, the probability that i and j are successive values in the chain, that is, the probability that the block ij appears in the partition. The relevances may be computed from the cohesions by a simple recursion requiring $O(n^2)$ elementary operations, given in a special case by Yao (1984).

Now suppose, that given the partition $i_0, \ldots, i_n$, the observations in different blocks are independent: there is a probability density f_{ij} associated with the block $i + 1, \ldots, j$ such that

$$f(X_1, \ldots, X_n | i_0, \ldots, i_n) = f_{0i_1}(X_1, \ldots, X_{i_1}) \cdots f_{i_{b-1}i_b}(X_{i_{b-1}+1}, \ldots, X_{i_b}).$$

Any joint distribution on observations and partitions that satisfies the product condition for partitions and the independence condition for observations given the partition will be called a product partition *model*. Let X_{ij} denote the observations $X_{i+1}, \ldots, X_j$. The element $f_{ij}(X_{ij})$ is called the *data factor*.

Given the observations $X_1, \ldots, X_n$, the partition $i_0, \ldots, i_n$ still has a product partition distribution, with posterior probabilities $p(\rho | \mathbf{X})$ determined by the posterior cohesions $f_{ij}(X_{ij})p_{ij}$. There will be a set of posterior transition probabilities that may be computed from the posterior cohesions. 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 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; such circumstances will occur if there are sharp changes in the distributions across change points, or if there are long sequences of data values in each subsequence.

In this paper we will investigate conditions under which the observations $X_1, X_2, \ldots, X_n$ do eventually dominate the prior cohesions c_{ij}, so that the true partition can be eventually approximately identified. These results suggest that the cohesions be chosen so that $c_{ij} = (j - i)^k$ for internal blocks and

$c_{0i} = c_{n-in} = i^{k+1}$, where $k < -2$. Such cohesions encourage the formation of large blocks, so that if there are no change points, we are likely to draw that conclusion.

2. Clustered parameters. A simple way to generate product partition models is through clustered parameters: a sequence of parameter values $\theta_1, \ldots, \theta_n$ ranging over some space Θ is partitioned into b contiguous subsets (or *blocks*) where the values are constant, that is, there exists a partition $\rho = (i_0, i_1, \ldots, i_n)$ of the set $\{1, 2, \ldots, n\}$ such that

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

$$n = i_b = i_{b+1} = \cdots = i_n$$

and

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

for $r = 1, 2, \ldots, b$. The parameter values change after the points $i_1, i_2, \ldots, i_{b-1}$. The value of the parameter in the block ij is θ_{ij}.

We construct a prior distribution for $\theta_1, \theta_2, \ldots, \theta_n$ as follows.

(a) The prior probability of $\rho = (i_0, i_1, \ldots, i_n)$ is

$$p(\rho) = \prod_{r=1}^{b} p_{i_{r-1} i_r},$$

where the quantity p_{ij} is the probability of a transition of endpoints from i to j.

(b) Given a partition ρ with b components, $\theta_{i_0 i_1}, \theta_{i_1 i_2}, \ldots, \theta_{i_{b-1} i_b}$ are independent, with θ_{ij} having density $f_{ij}(\theta_{ij})$ with respect to some measure on Θ (which we will represent in integrals as $d\theta$).

The joint distribution on partitions and parameters is a product partition model according to the definition of Section 1. In this case, when we know the parameters, we know the partition exactly.

The observations $X_1, \ldots, X_n$ are assumed independent given $\theta_1, \ldots, \theta_n$ having joint density $\prod f(X_i|\theta_i)$. The joint density of observations and parameters given ρ is a product of densities over the different blocks in ρ, with the density in block ij being

$$f_{ij}(X_{ij}, \theta_{ij}) = \prod_{k=i+1}^{j} f(X_k|\theta_{ij}) f_{ij}(\theta_{ij}).$$

Thus the joint distribution of the partition, the parameters and the observations forms a product partition model. The joint distribution of the partition and observations is a product partition model with data factor

$$f_{ij}(X_{ij}) = \int \left[\prod_{k=i+1}^{j} f(X_k|\theta) \right] f_{ij}(\theta)\, d\theta.$$

The conditional distribution of partition and parameters given the observations is also a product partition model with cohesions $p_{ij} f_{ij}(X_{ij})$ and data

factor the posterior density

$$f_{ij}(\theta_{ij}|X_{ij}) = f_{ij}(\theta_{ij})\left[\prod_{k=i+1}^{j} f(X_k|\theta_{ij})\right] \Big/ f_{ij}(X_{ij}).$$

The product partition model thus offers a smooth machinery for handling inferences about clustered parameters. We do the standard Bayes calculations within each possible block ij, computing the posterior distribution of the parameter values for each block using the observations in the block. We make inferences about the unknown partition using the prior cohesions and the data factor for each block which is equal to the marginal density of observations in the block. And then we combine these two types of information in a final product model conditional on the observations.

A typical calculation for the combined model would be the conditional density of θ_k given the observations

$$f(\theta_k|\mathbf{X}) = \sum_{i<k\leq j} f_{ij}(\theta_k|X_{ij})r(ij|\mathbf{X}),$$

where $r(ij|\mathbf{X})$ is the posterior relevance of ij, the probability that the block ij appears in the partition, given the data $\mathbf{X}$. Thus in making inferences about θ_k, we will allow the data to adapt to the true partition; if it appears that observations in a block about k are from the same density, then all those observations will be used in making an inference about θ_k; and the final density will be an average of densities computed for blocks including k and weighted by the probabilities that those blocks are the correct ones for making inferences about θ_k.

We will consider in some detail a change point model introduced by Duncan (1956) and intensively developed by Yao (1984). Yao gives the recursive formula for computing relevances from cohesions, which permits $O(n^2)$ computations of posterior expectations.

In this model, observations X_i are independently normal with means μ_i and variance σ^2. The means μ_i change after points $i_1, i_2, \ldots, i_{b-1}$; the probability that $\mu_{i+1} = \mu_i$ given all past means is $1 - p$. The points $i_1, i_2, \ldots, i_{b-1}$ define a partition of the observations composed of the sets $0i_1, i_1i_2, \ldots, i_{b-1}n$ with cohesions $(1-p)^{i_1-1}p, (1-p)^{i_2-i_1-1}p, \ldots, (1-p)^{n-i_{b-1}-1}$. Second, the different means μ_i are independently sampled from $N(\mu_0, \sigma_0^2)$; thus when the partition ρ is known, the observations in different components are independent, after averaging over the possible mean values μ_i in each partition. The probabilities $f_{ij}(X_{ij})$ of the observations $X_{i+1}, \ldots, X_j$ thus conform to a product partition model. A slightly different model is considered by Barnard (1959) and Chernoff and Zacks (1964); in that model, the mean μ_i equals the mean μ_{i-1} plus a random increment. This is certainly a plausible mechanism, but the independence of observations in different blocks is lost and the simple calculations of the product model are no longer possible.

We consider the class of cohesions that ensure the final distribution of $X_1, \ldots, X_n$ is stationary; the geometric distribution studied by Yao is a fruitful

member of this class, but we propose a long-tailed polynomial distribution that makes larger intervals of constant means relatively more probable. Using this distribution, we show that if observations are actually sampled from a constant mean, the posterior probability of the event $\mu_1 = \mu_2 = \cdots = \mu_n$ approaches 1. And if the observations are actually sampled from a partition with one change point, the posterior probability of partitions having components close to the true components approaches 1. We speculate that the same result will hold for any number of change points; thus preliminary evidence suggests that the polynomial based cohesions give consistent inferences, successfully approximating the true partition asymptotically.

3. Some examples.

EXAMPLE 1 (Normals). This is the case studied in Yao (1984). Let $f(X_i; \theta_i)$ be the normal density with mean θ_i and variance σ^2. Let the prior density of θ_{ij} be normal with mean μ_0 and variance σ_0^2.

Then

$$f_{ij}(X_{ij}) = \frac{1}{(2\pi\sigma^2)^{(j-i)/2}} \left(\frac{\sigma^2}{(j-i)\sigma_0^2 + \sigma^2} \right)^{1/2} \times \exp\left\{ - \frac{\sum_{l=i+1}^{j} (X_l - \mu_0)^2}{2\sigma^2} + \frac{\sigma_0^2 \left[\sum_{l=i+1}^{j} (X_l - \mu_0) \right]^2}{2\sigma^2 \left((j-i)\sigma_0^2 + \sigma^2 \right)} \right\}.$$

The parameters $\mu_0, \sigma^2, \sigma_0^2$ must be estimated from the data with some care; the EM algorithm [Dempster, Laird and Rubin (1977)] is helpful if the parameters are estimated by maximum likelihood. A more difficult question is the choice of the prior partition distribution, which we will return to in Sections 6 and 7.

EXAMPLE 2 (Binomials). Let

$$f(X; \theta_i) = \begin{cases} \theta_i, & \text{if } X = 1, \\ 1 - \theta_i, & \text{if } X = 0. \end{cases}$$

Let θ_{ij} have a beta prior density with parameters m_1, m_2. Then

$$f_{ij}(X_{ij}) = \frac{\beta\left(m_1 + \sum_{l=i+1}^{j} X_l, m_2 + j - i - \sum_{l=i+1}^{j} X_l \right)}{\beta(m_1, m_2)}.$$

EXAMPLE 3 (Regressions). Let $\{Y_i\colon 1 \le i \le n\}$ and $\{X_i\colon 1 \le i \le n\}$ be two time series and suppose we are interested in the regression of Y on X. Assume

$$Y_i = \alpha_i + \beta_i X_i + e_i, \qquad i = 1, 2, \ldots, n,$$

where the errors $\{e_i\}$ are i.i.d. $N(0, \sigma^2)$.

Here $\theta_i = (\alpha_i, \beta_i)$ and a convenient choice for the prior density of θ_{ij} is the bivariate normal density. See Raiffa and Schlaifer (1961), Chapter 13, for the technical details necessary to calculate $f_{ij}(X_{ij})$ and $E[\theta_i|\mathbf{X}]$.

Extensions to multiple regression change point problems are equally staightforward.

EXAMPLE 4 (Histograms). Let $X_1, \ldots, X_n$ be a series of observations in the unit interval. We wish to construct a histogram for these observations. Traditionally, we partition the interval $(0, 1)$ into m intervals of equal size and take the density in each of these intervals to be proportional to the number of observations in the interval. But there may be benefits in allowing intervals of different sizes in different parts of the data—large intervals when the underlying unknown population density is small or changes slowly, small intervals when it is large or changes rapidly.

In a product partition approach, we construct a distribution over the set of possible histograms as follows:

1. The intervals of the histograms are the intervals $I_1, I_2, \ldots, I_k$ with probability proportional to $\prod c(I_j)$, where $c(I_j)$ is the cohesion associated with interval I_j. For simplicity, the intervals I_j will be assumed to have endpoints on some discrete grid on $(0, 1)$. Frequently the only possible values of observations fall on such a grid.
2. If there are k intervals $I_1, I_2, \ldots, I_k$ in the histogram, the probabilities $p_1, p_2, \ldots, p_k$ assigned to the intervals are sampled from a Dirichlet distribution with parameters $\alpha(I_1), \alpha(I_2), \ldots, \alpha(I_k)$, where α is a measure on $(0, 1)$. Thus $p_1, \ldots, p_{k-1}$ has density

$$\Gamma[\alpha(0,1)] \prod_j \frac{p_j^{\alpha(I_j)-1}}{\Gamma[\alpha(I_j)]}.$$

The probability density of the observations $X_1, \ldots, X_n$ given $\{I_j\}$ and $\{p_j\}$ is

$$\prod_j \left(\frac{p_j}{|I_j|}\right)^{n(I_j)}$$

where $n(I_j)$ is the number of observations in the interval I_j. The probability density of the observations given $\{I_j\}$ alone is

$$\frac{\Gamma[\alpha(0,1)]}{\Gamma[\alpha(0,1)+n]} \prod_j \frac{\Gamma[\alpha(I_j)+n(I_j)]}{\Gamma[\alpha(I_j)]|I_j|^{n(I_j)}}.$$

The distribution on $\{I_j\}$ given the observations is thus a product distribution with posterior cohesions

$$\frac{c(I_j)\Gamma[\alpha(I_j)+n(I_j)]}{|I_j|^{n(I_j)}\Gamma[\alpha(I_j)]}.$$

It is computationally feasible to compute the posterior probability that an interval I lies in the partition. The estimated density at a point x is

$$\hat{f}(x) = \sum_{I|x\in I} \frac{\alpha(I) + n(I)}{(\alpha(0,1) + n)} \frac{1}{|I|} P(I \in \rho|\mathbf{X}).$$

Thus if the data suggest that long intervals I near x are appropriate, the density estimate will be close to the histogram estimate based on long intervals. If the data suggest short intervals are appropriate, the density estimate will be near the histogram estimate based on short intervals.

A plausible measure α is the uniform on $(0, 1)$ with total weight 1. We would like to select the prior cohesions to discourage partitions with many intervals. For example, $c(I) = \lambda/N$, where interval endpoints are of form i/N, $i = 0, 1, 2, \ldots, N$. The expected number of intervals in a partition is then $1 + \lambda$; and the probability of k intervals in a partition is approximately $(\lambda^{k-1}/(k-1)!)e^{-\lambda}$. We might select λ by maximum likelihood, or by beginning with a prior distribution on λ uniform over $\lambda = 1, 2, 5, 10, 20$. Computations would be done for each of these 5 values of λ and averaged with respect to the posterior probabilities of λ given the data.

Finally, an optimal histogram would be the partition of optimal posterior probability.

4. 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 $\lambda(r, s) = \Sigma\Pi_{j=1}^{b} c_{i_{j-1}i_j}$, where the summation is over all sets of integers $r = i_0 < i_1 < \cdots < i_b = s$. Then

$$r(ij) = \frac{\lambda(0, i)c_{ij}\lambda(j, n)}{\lambda(0, n)}.$$

The quantities $\lambda(0, r)$ and $\lambda(r, n)$ may be calculated in $O(n^2)$ steps using the recursions

$$\lambda(0, 1) = c_{01},$$

$$\lambda(0, r+1) = c_{0,r+1} + \sum_{t=1}^{r} \lambda(0, t)c_{t,r+1}$$

and

$$\lambda(n-1, n) = c_{n-1,n},$$

$$\lambda(r, n) = c_{rn} + \sum_{t=r+1}^{n-1} c_{rt}\lambda(t, n).$$

The posterior relevances are computed from posterior cohesions by the same recursions. For each i there are $O(n^2)$ sets jk which contain i and so

$\{E(\theta_i|\mathbf{X})\colon i = 1, 2, \ldots, n\}$ may be calculated in $O(n^3)$ steps. The above recursions are given by Yao (1984).

Since 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 formulae. These computations are useful in estimating the various nuisance parameters in the model.

5. Stationary product partition distributions. A product partition distribution is *stationary* if the probability that $i + 1, \ldots, j$ all lie in the same block in the random partition ρ depends only on $j - i$. Equivalently,

$$\sum_{r=0}^{n-1} P\{i_r \le i, i_{r+1} \ge j\} \quad \text{depends only on } j - i.$$

LEMMA 1. *Let J be an arbitrary random variable on the positive integers, with finite average PJ. Define I_1 by $P\{I_1 = k\} = P\{J \ge k\}/PJ$. Let $J_1, J_2, \ldots$ be independent realizations of J. Let $X \wedge n$ be $\min(X, n)$. Then $0, I_1 \wedge n, (I_1 + J_1) \wedge n, (I_1 + J_1 + J_2) \wedge n, \ldots$ forms a stationary product partition distribution.*

The proof will be omitted. The variable J is called a *jump* variable. The partition endpoints may be constructed from a renewal process $0, J_1, J_1 + J_2, \ldots$ as follows: shift the process by subtracting the integer N. Let I_1^N be the first value in the shifted process that lies in the interval $[1, n]$ and define $I_2^N, I_3^N, \ldots$ to be subsequent values of the renewal process that lie in $[1, n]$. Now let $N \to \infty$. The limiting distribution of the I_r^N is the distribution of $I_1 \wedge n, (I_1 + J_1) \wedge n, (I_1 + J_1 + J_2) \wedge n, \ldots$.

The transition to j given i has the distribution of $J + i$, with the modification that all values of j greater than n are reset to take the value n.

Let $g(j) = P\{J = j\}$ for each positive integer j. Let $G(j) = \sum_{i=j}^{\infty} g(i)$. Let $G_0 = \sum jg(j)$.

In this case, the transition probabilities for the change points are:

$$\begin{aligned}
p_{0n} &= \sum_{j \ge n} G(j)/G_0, \\
p_{0j} &= G(j)/G_0, \qquad 1 \le j < n, \\
p_{in} &= G(n - i), \qquad 0 < i < n, \\
p_{ij} &= g(j - i), \qquad 0 < i < j < n.
\end{aligned}$$

Say that the function g has *tail power* k if $g(j)j^{-k}$ converges to some nonzero limit as $j \to \infty$. A function with tail power k behaves similarly to a slowly varying function with exponent k. If g has tail power $k < -1$, then G has tail power $k + 1$.

For example, let J take value j, $1 \le j < \infty$, with probability $g(j) = 4/(j(j+1)(j+2))$ having tail power -3. Then $PJ = 2$ and I_1 takes the value i, $1 \le i < \infty$, with probability $1/(i(i+1))$. The transition probabilities are:

$$\begin{aligned} p_{0n} &= 1/n, \\ p_{0j} &= 1/(j(j+1)), && 1 \le j < n, \\ p_{in} &= 2/((n-i)(n-i+1)), && 0 < i, \\ p_{ij} &= 4/((j-i)(j-i+1)(j-i+2)), && 0 < i < j < n. \end{aligned}$$

Note that the transition probabilities p_{ij} that do not involve endpoints $0, n$ depend only on $j - i$, but that the endpoint transitions are somewhat different. We imagine the renewal process starting in the remote past, hitting the interval $1, \ldots, n$ at some point I_1 for the first time, proceeding with stationary transitions through $i_2, \ldots, i_{b-1}$, then exiting beyond n; neither 0 nor n are necessarily points of the renewal process, so the transition probabilities at those points are not the same as the transition probabilities for interior points. Although the transition probabilities p_{0j} and $p_{n-j,n}$ differ by a constant multiple, they make the same effective contribution to product probabilities of partitions.

6. Consistency when there are no change points. Let ρ_0 denote the partition consisting of the single block $0n$. The density of the observations when this partition is true is $f(\mathbf{X}|\rho_0) = f_{0n}(\mathbf{X}_{0n})$. The change point model is *consistent* for ρ_0 if, when the observations are sampled according to f_{0n}, $p(\rho_0|\mathbf{X}_{0n}) \to 1$ in probability as $n \to \infty$.

Suppose that the prior distribution on ρ is stationary with jump variable J having density g. The following lemma will be used to bound the probabilities that many jumps occur given the data. In the statement and proof of the lemma, we use the same symbol C for the constant in inequalities, even though different constants may be appropriate for the different inequalities.

LEMMA 2. *Let g be nonnegative with tail power $k < -2$. Let $\alpha \in [2/3, 1)$. There exists a constant C such that*

(a) $$\sum_{i=1}^{n-1} \left(\frac{G(i)}{\sqrt{i}}\right)^{\alpha} \left(\frac{G(n-i)}{\sqrt{n-i}}\right)^{\alpha} \le C\left(\frac{G(n)}{\sqrt{n}}\right)^{\alpha}$$

for all $n > 1$;

(b) $$\sum_{i=j}^{n-1} \left(\frac{g(i)}{\sqrt{i}}\right)^{\alpha} \left(\frac{G(n-i)}{\sqrt{n-i}}\right)^{\alpha} \le C\left(\frac{G(n)}{j\sqrt{n}}\right)^{\alpha}$$

for all $1 \le j < n < \infty$;

$$\text{(c)} \quad \sum_{r_1+\cdots+r_b=n} \prod_{j=1}^{b} \left(\frac{g(r_j)}{\sqrt{r_j}}\right)^{\alpha} \le Cb^{1+((1/2)-k)\alpha}\left[\sum_{j=1}^{\infty}\left(\frac{g(j)}{\sqrt{j}}\right)^{\alpha}\right]^{b-1}\left(\frac{g(n)}{\sqrt{n}}\right)^{\alpha}$$

for all b, n.

PROOF. We prove (a) and (b) by replacing $g(i)$ by i^k and $G(i)$ by i^{k+1}. Then $(g(i)/\sqrt{i})^{\alpha} = i^{\beta}$, where $\beta = (k - 1/2)\alpha < -1$ and the results follow from standard manipulations.

To prove (c), we first consider those partitions $n = r_1 + \cdots + r_b$ for which the maximal r_i is r_b (the overall sum will be less than the sum of contributions where successively $r_1, r_2, \ldots, r_b$ are maximal and so it is less than b times the contribution from the partitions where r_b is maximal). Necessarily $r_b \ge n/b$ and $g(i)i^{-k}$ converges to some nonzero limit, so

$$\frac{g(r_b)}{\sqrt{r_b}} \le C\frac{g(n)}{\sqrt{n}}b^{(1/2)-k},$$

$$\sum_{r_1+\cdots+r_b=n} \prod_{j=1}^{b} \left(\frac{g(r_j)}{\sqrt{r_j}}\right)^{\alpha} \le Cb \sum_{r_1+\cdots+r_{b-1}\le n} \prod_{j=1}^{b-1} \left(\frac{g(r_j)}{\sqrt{r_j}}\right)^{\alpha} \left(\frac{g(n)}{\sqrt{n}}b^{(1/2)-k}\right)^{\alpha}$$

$$\le Cb \sum_{r_1>0,\ldots,r_{b-1}>0} \prod_{j=1}^{b-1} \left(\frac{g(r_j)}{\sqrt{r_j}}\right)^{\alpha} \left(\frac{g(n)}{\sqrt{n}}b^{(1/2)-k}\right)^{\alpha}$$

$$\le Cb^{1+((1/2)-k)\alpha}\left[\sum_{j=1}^{\infty}\left(\frac{g(j)}{\sqrt{j}}\right)^{\alpha}\right]^{b-1}\left(\frac{g(n)}{\sqrt{n}}\right)^{\alpha}. \qquad \square$$

THEOREM 1. *Let* P_0 *be some null distribution for* $X_1, \ldots, X_n, \ldots$. *Suppose that the prior distribution of the partition* ρ *is stationary with jump variable* J *with density* g *having tail power* $k < -2$. *Suppose that for some fixed* $\alpha \in [2/3, 1)$, δ *such that* $\delta\Sigma(g(j)/\sqrt{j})^{\alpha} < 1$ *and fixed* C,

$$\text{(1)} \qquad P_0\left[\frac{f(\mathbf{X}|\rho)}{f(\mathbf{X}|\rho_0)}\right]^{\alpha} \le C\delta^{b(\rho)}\left(\frac{\sqrt{n}}{\prod_{ij\in\rho}\sqrt{(j-i)}}\right)^{\alpha}$$

for all ρ. *Then*

$$p(\rho_0|\mathbf{X}) \to 1 \quad \textit{in } P_0\textit{-probability}.$$

PROOF. The notation PY denotes the expectation of Y with respect to the probability measure P.

We will first prove that the probability of more than two change points is of the same order of magnitude asymptotically as precisely two change points. We

then prove that either one or two change points have negligible probability compared to zero change points, so that zero change points occur asymptotically in probability.

Recall that $b = b(\rho)$ is the number of blocks in a partition ρ having change points $i_1, \ldots, i_{b-1}$. Suppose $b > 2$.

Using inequality (1) and setting

$$T_{i_1 i_{b-1}} = \frac{p_{0i_1} g(i_{b-1} - i_1) p_{i_{b-1}n} / p_{0n}}{\left(i_1(i_{b-1} - i_1)(n - i_{b-1})/n\right)^{1/2}},$$

$$\frac{p(\rho|\mathbf{X})}{p(\rho_0|\mathbf{X})} = \frac{f(\mathbf{X}|\rho)p(\rho)}{f(\mathbf{X}|\rho_0)p(\rho_0)},$$

$$P_0\left(\frac{p(\rho|\mathbf{X})}{p(\rho_0|\mathbf{X})}\right)^{\alpha} \le \left(T_{i_1 i_{b-1}}\right)^{\alpha} C\delta^b \left[\frac{\prod_{j=2}^{b-1} g(i_j - \hat{i}_{j-1}) / \sqrt{(i_j - i_{j-1})}}{g(i_{b-1} - i_1) / \sqrt{(i_{b-1} - i_1)}}\right]^{\alpha}.$$

Now consider all partitions ρ' with a given first change point i_1 and final change point i_{b-1}. Using Lemma 2(c) and the fact that $(\sum u_i)^{\alpha} \le (\sum u_i^{\alpha})$ for $u_i \ge 0$ and $0 < \alpha \le 1$,

$$P_0\left(\sum \frac{p(\rho'|\mathbf{X})}{p(\rho_0|\mathbf{X})}\right)^{\alpha} \le \left(T_{i_1 i_{b-1}}\right)^{\alpha} C \sum_{b>2} \delta^b b^{1+((1/2)-k)\alpha} \left[\sum_{j=1}^{\infty} \left(\frac{g(j)}{\sqrt{j}}\right)^{\alpha}\right]^{b-3}$$

$$\le T_{i_1 i_{b-1}}^{\alpha} \Sigma_{\delta},$$

where $\Sigma_{\delta} < \infty$ since $\delta \sum_{j=1}^{\infty} (g(j)/\sqrt{j})^{\alpha} < 1$. Because the power series converges, all partitions of i_1, i_{b-1} together have the same order of probability as the single block i_1, i_{b-1}.

We now collect all partitions ρ'' with $b > 2$ by summing over i_1, i_{b-1}.

$$P_0\left(\sum \frac{p(\rho''|\mathbf{X})}{p(\rho_0|\mathbf{X})}\right)^{\alpha} \le \Sigma_{\delta} \sum T_{i_1 i_{b-1}}^{a}$$

$$= \Sigma_{\delta} \sum \left[\frac{G(i_1) g(i_{b-1} - i_1) G(n - i_{b-1}) / (p_{0n} G_0)}{\left(i_1(i_{b-1} - i_1)(n - i_{b-1})/n\right)^{1/2}}\right]^{\alpha}$$

$$\le \Sigma_{\delta} C \left(\frac{G(n)}{p_{0n} G_0}\right)^{\alpha} \quad \text{[by Lemma 2 (a), (b)]}.$$

Since G has tail power $k + 1$ and $k < -2$, $p_{0n} = \sum_{j \ge n} G(j)/G_0$ has tail power $k + 2$, so that $G(n)/p_{0n} \to 0$ as $n \to \infty$. Thus $p\{b(\rho) > 2|\mathbf{X}\}$ approaches zero in P_0-probability. The case $b = 2$ is handled similarly, with the simplification that the product terms due to partitioning i_1, i_{b-1} disappear. Thus $p\{b(\rho) \ge 2|\mathbf{X}\}$ approaches zero in P_0-probability, as required. □

We need to demonstrate that inequality (1) is satisfied in interesting cases. Suppose that the clustered parameter model applies for r-dimensional parame-

ters $\theta_1, \ldots, \theta_n$; the block parameters θ_{ij} are assumed to have prior density $f(\theta)$. Under suitable regularity conditions, if the observations $X_1, \ldots, X_n$ are sampled from the density $f(X|\theta_0)$, the marginal density $f_{0n}(X_{0n})$ is approximately, for large n,

$$f(\theta_0)n^{-1/2}2\pi^{r/2}|I(\theta_0)|^{-1/2}\exp\left[\tfrac{1}{2}n(\hat{\theta}_n - \theta_0)'I(\theta_0)(\hat{\theta}_n - \theta_0)\right]\prod f(X_i|\theta_0),$$

where $|I(\theta_0)|$ is the determinant of the information matrix $I(\theta_0)$ at θ_0 and $\hat{\theta}_0$ is the maximum likelihood estimate [see Hartigan (1983), page 108, for example]. The exponential component is distributed approximately as $e^{\chi_r^2/2}$; the average value of $e^{\alpha\chi_r^2/2}$ is $(1-\alpha)^{-r/2}$.

Taking this approximation to be exact, with P_0 denoting sampling from the density $f(X|\theta_0)$,

$$P_0\left[\frac{f(\mathbf{X}|\rho)}{f(\mathbf{X}|\rho_0)}\right]^{\alpha} = \left[f(\theta_0)2\pi^{r/2}|I(\theta_0)|^{-1/2}\right]^{\alpha(b-1)} \times\left(\frac{\sqrt{n}}{\prod_{ij\in\rho}\sqrt{(j-i)}}\right)^{\alpha}(1-\alpha)^{-r(b-1)/2}.$$

Thus the inequality assumed in the theorem is met with

$$\delta = \left[f(\theta_0)2\pi^{r/2}|I(\theta_0)|^{-1/2}\right]^{\alpha}(1-\alpha)^{-r/2},$$

provided that $\delta\Sigma(g(j)/\sqrt{j})^{\alpha} < 1$. If $f(\theta_0)$ is chosen small enough (so that the prior density is not too highly concentrated at θ_0), the inequality (1) will be satisfied.

The more natural null distribution is that according to ρ_0, which is achieved by selecting θ_0 at random from the density f and then selecting the sequence $X_1, \ldots, X_n, \ldots$ from $f(X|\theta_0)$. In this case, when the approximation is exact, we set

$$\delta = \sup_{\theta_0}\left[f(\theta_0)2\pi^{r/2}|I(\theta_0)|^{-1/2}\right]^{\alpha}(1-\alpha)^{-r/2}.$$

It will be possible to have condition (1) satisfied provided that $f(\theta_0)2\pi^{r/2}|I(\theta_0)|^{-1/2}$ is everywhere small, which means that the prior density is everywhere small compared to Jeffreys' density.

We have made the argument that the inequality (1) should apply quite widely in clustered parameter models, provided the prior distribution for the parameter is sufficiently diffuse. Let us consider the Duncan model, where the observations are independent normal with mean θ and variance 1 and the prior distribution for the parameter θ is normal with mean 0 and variance σ_0^2. We will use the fact that

$$\frac{n+a}{(n_1+a)(n_2+a)} \leq \frac{n}{n_1 n_2} \quad \text{for } a \geq 0,\ n \geq n_1, n_2.$$

Let $\hat{X}_{ij} = \Sigma_{k \in ij} X_k/(j-i)$:

$$f_{0n}(X_{0n}) = (2\pi)^{-n/2}\left(1 + n\sigma_0^2\right)^{-1/2} \exp\left[-\frac{1}{2}\sum X_i^2 + \frac{n\hat{X}_{0n}^2/2}{1 + 1/\left(n\sigma_0^2\right)}\right],$$

$$\frac{f(\mathbf{X}|\rho)}{f(\mathbf{X}|\rho_0)} \le \sigma_0^{1-b} \exp\left[-\frac{n\hat{X}_{0n}^2/2}{1 + 1/\left(n\sigma_0^2\right)} + \sum_{ij \in \rho} \frac{(j-i)\hat{X}_{ij}^2/2}{1 + 1/\left((j-i)\sigma_0^2\right)}\right]$$

$$\times \frac{\sqrt{n}}{\Pi_{ij \in \rho}\sqrt{j-i}},$$

$$\frac{-n\hat{X}_{0n}^2}{1 + 1/\left(n\sigma_0^2\right)} + \sum_{ij \in \rho} \frac{(j-i)\hat{X}_{ij}^2}{1 + 1/\left((j-i)\sigma_0^2\right)} \le -n\hat{X}_{0n}^2 + \sum_{ij \in \rho} (j-i)\hat{X}_{ij}^2 \sim \chi_{b-1}^2.$$

Thus

$$P_0\left[\frac{f(\mathbf{X}|\rho)}{f(\mathbf{X}|\rho_0)}\right]^{\alpha} \le \sigma_0^{(1-b)\alpha}(1-\alpha)^{-(b-1)/2}\left(\frac{\sqrt{n}}{\Pi_{ij \in \rho}\sqrt{j-i}}\right)^{\alpha},$$

and inequality (1) is satisfied with $\delta = \sigma_0^{-\alpha}(1-\alpha)^{-1/2}$. For any prior distribution with tail power k, if σ_0^2 is large enough, then consistency will be achieved when sampling from a normal with mean θ_0 and variance 1, or when sampling from $f_{0n}(X_{0n})$, since the distribution of $-n\hat{X}_{0n}^2 + \Sigma_{ij \in \rho}(j-i)\hat{X}_{ij}^2$ is χ_{b-1}^2 in either case.

For example, if $g(i) = 4/(i(i+1)(i+2))$ and $\alpha = 2/3$, then $\Sigma(g(i)/\sqrt{i})^{\alpha} = 1.39$ and $\sigma_0 > 3.74$ ensures consistency. Consistency may be achieved for lower values of σ_0, indeed perhaps for all σ_0.

7. Consistency for one change point. If the true partition contains several change points, we cannot achieve consistency in the strong sense that the distribution over partitions concentrates on the true partition, but only in the weaker sense that each random change point is within $O_p(1)$ of some true change point. The following theorem makes this assertion precise for a single change point. In the proof, we will use C for a generic constant that may take different values in different inequalities.

THEOREM 2. *Let the prior distribution of change points be stationary with jump variable having density g with tail power $k < -2$. Suppose that the true partition ρ_m has a single change point at m, where $m/\log(n) \to \infty$, $(n-m)/\log(n) \to \infty$. For any partition ρ, let $i_r i_{r+1}$ be the block of ρ that includes m.*

Let P_n be a sequence of probability distributions, consistent with a single change point: For each $\varepsilon_0 > 0$, there exists a sequence of 0–1 variables Z_n such that $P_n Z_n > 1 - \varepsilon_0$, and for some fixed $\alpha \in (2/3, 1)$, δ such that $\delta \Sigma (g(j)/\sqrt{j})^\alpha < 1$ and fixed C, $0 < \Delta < 1$,

$$(2) \quad P_n\left[\frac{Z_n f(\mathbf{X}|\rho)}{f(\mathbf{X}|\rho_m)}\right]^\alpha \le C\delta^{b(\rho)}\left[\frac{m(n-m)}{\Pi_{ij\in\rho}(j-i)}\right]^{\alpha/2} \Delta^{1/((1/(m-i_r))+(1/(i_{r+1}-m)))}.$$

Then $P_n[p\{\max_j |i_j - m| > A|\mathbf{X}\} > \varepsilon] \to 0$ as $A, n \to \infty$ for each $\varepsilon > 0$.

PROOF. The theorem asserts that all the random change points will be within $O_p(1)$ of the true change point, according to the posterior distribution given $\mathbf{X}$, for observations $\mathbf{X}$ taken according to P_n. An obvious choice of P_n is just the true change point distribution with density $f_{0m}(X_{0m})f_{mn}(X_{mn})$, but other choices are plausible in parametric problems, where we might wish to assume fixed different parameter values in the two true blocks. The sequence Z_n weakens the condition to accommodate cases where the inequality applies once a small set of bad observations is excluded. To simplify the proof we will assume the Z_n are all equal to 1; the more general proof requires some small extra manipulations based on the inequality $P_n\{\mathbf{X} \in B\} \le P_n Z_n\{\mathbf{X} \in B\} + \varepsilon_0$. The condition (2) differs from the similar condition (1) in the no-change point case in the term $\Delta^{1/((1/(m-i_r))+(1/(i_{r+1}-m)))}$ that is generated by the block in ρ that includes m; the effect of this term is to inhibit the formation of such blocks with boundaries far from m.

We will use the measure on partitions ρ of $0n$ defined by

$$E_n B = \sum_{\rho\in B} \delta^{b(\rho)-1}\left[\frac{\Pi_{ij\in\rho}\left(p_{ij}/\sqrt{j-i}\right)}{\left(p_{0m}/\sqrt{m}\right)\left(p_{mn}/\sqrt{n-m}\right)}\right]^\alpha \Delta^{1/((1/(m-i_r))+(i/(i_{r+1}-m)))}.$$

Since, by condition (2), and using $p(\rho_m|\mathbf{X}) \le 1$,

$$P_n p^\alpha\{\rho \in B|\mathbf{X}\} \le P_n \sum_{\rho\in B}\left[\frac{p(\rho|\mathbf{X})}{p(\rho_m|\mathbf{X})}\right]^\alpha \le CE_n B,$$

whenever $E_n B \to 0$ as $n \to \infty$, then $p\{\rho \in B|\mathbf{X}\} \to 0$ in P_n-probability.

Consider the sets of partitions

$$B_{ij} = \{p|ij \text{ is a block in } \rho\},$$

$$B^*_{ij} = \{\rho|i, j \text{ are change points in } \rho\}.$$

The set B^*_{ij} includes B_{ij} but includes in addition those partitions in which change points occur between i and j. It follows from Lemma 2(c), by summing

over all partitions of ij, that if $i, j \le m$ or $i, j \ge m$, then $E_n B^*_{ij} \le CE_n B_{ij}$. Thus a set B^*_{ij} is asymptotically negligible whenever B_{ij} is asymptotically negligible.

We need to show that $E_n\{|i_1 - m| > A\} \to 0$ as n, A approach ∞. (By symmetry, this result will be true for i_{b-1} if it is true for i_1 and therefore will be true for all change points.) We have seen that cases with more than two change points in $0m$ or in mn may be reduced to cases with two or fewer change points in $0m$ or in mn, after replacing sets B^*_{ij} with sets B_{ij}. We need to consider only the cases of (i) zero change points, (ii) one change point in $0m$, (iii) two change points, both in $0m$, (iv) two change points, one on either side of m, (v) three change points of which two lie in $0m$ and one in mn.

For example, the case of three change points in which one lies in $0m$ and two in mn reduces to the case of two change points, one on either side of m, since $B^*_{i_2 n}$ may be replaced by $B_{i_2 n}$. We cannot dismiss case (v) in the same way, because if $B^*_{0i_2}$ is replaced by B_{0i_2}, we can draw a conclusion only about i_2, not about our target point i_1.

Let $\beta = (k + 1/2)\alpha < -1$.

(i) Zero change points: Let ρ_0 denote the partition with no change point. Noting that $p_{0n} \sim n^{k+2}$, $p_{0i} \sim i^{k+1}$, $p_{in} \sim (n - i)^{k+1}$,

$$E_n \rho_0 \le C \frac{n^{\beta+\alpha}}{m^\beta (n-m)^\beta} \Delta^{m(n-m)/n}.$$

Since $m/\log(n), (n - m)/\log(n) \to \infty$, the exponential term will dominate the polynomial term as $n \to \infty$ and so $E_n \rho_0 \to 0$ as $n \to \infty$.

(ii) One change point, $i_1 \le m$: Let ρ_{i_1} be the partition with a single change point at i_1:

$$E_n \rho_{i_1} \le C \left[\frac{i_1(n - i_1)}{m(n-m)} \right]^\beta \Delta^{(n-m)(m-i_1)/(n-i_1)}.$$

Note that

$$\Delta^{(n-m)(m-i_1)/(n-i_1)} < \Delta^{(m-i_1)/2} + \Delta^{(n-m)/2},$$

$$\left[\frac{i_1(n - i_1)}{m(n-m)} \right]^\beta \le (1 + m - i_1)^{-\beta}$$

$$\le n^{-\beta}.$$

Since $(n - m)/\log(n) \to \infty$, $\Delta^{(n-m)/2} n^{1-\beta} \to 0$. Also, since $\Delta < 1$,

$$\sum (1 + j)^{-\beta} \Delta^{j/2} < \infty.$$

Thus

$$\sum_{i_1<m-A}\left[\frac{i_1(n-i_1)}{m(n-m)}\right]^{\beta}\Delta^{(n-m)(m-i_1)/(n-i_1)}$$

$$< n^{1-\beta}\Delta^{(n-m)/2} + \sum_{i_1<m-A}(1+m-i_1)^{-\beta}\Delta^{(m-i_1)/2}$$

$$\to 0 \quad \text{as } A, n \to \infty.$$

Thus, when there is a single change point, $E_n\{|i_1 - m| > A\} \to 0$ as $A, n \to \infty$.

(iii) Two change points, $i_1, i_2 \le m$: Let $\rho_{i_1 i_2}$ denote the partition with change points at i_1 and i_2:

$$\sum_{i_1} E_n\rho_{i_1 i_2} \le CE_n\rho_{i_2} \quad \text{using Lemma 2(c)}.$$

Thus, by the argument in (ii),

$$E_n\{i_2 < m - A\} = \sum_{i_1, i_2 | i_2 < m-A} E_n\rho_{i_1 i_2} \to 0 \quad \text{as } A, n \to \infty.$$

Also

$$\sum_{i_1 | i_2 - i_1 > A} E_n\rho_{i_1 i_2} \le CE_n\rho_{i_2}A^{-\alpha} \quad \text{using Lemma 2(b)}.$$

Since $\Sigma E_n\rho_{i_2}$ is bounded, $E_n\{i_1 < i_2 - A\} \to 0$ as $A, n \to \infty$. Thus $E_n\{i_1 < m - 2A\} \to 0$ as $A, n \to \infty$.

(iv) Two change points, $i_1 \le m < i_2$: Assume without loss of generality that $i_1 + i_2 \le 2m$. (Otherwise, by symmetry, we could show that $E_n\{i_2 > m + A\} \to 0$ *as* $A, n \to \infty$.) Then

$$\Delta^{(m-i_1)(i_2-m)/(i_2-i_1)} \le \Delta^{(i_2-m)/2},$$

$$E_n\rho_{i_1 i_2} \le C\left[\frac{i_1(i_2-i_1)(n-i_2)}{m(n-m)}\right]^{\beta}(i_2-i_1)^{-\alpha}\Delta^{(m-i_1)(i_2-m)/(i_2-i_1)},$$

$$E_n\{i_1 < m - A\} \le \sum_{i_2 - i_1 > A} E_n\rho_{i_1 i_2}$$

$$\le \sum_{i_2} CA^{-\alpha}(1 + i_2 - m)^{-\beta}\Delta^{(i_2-m)/2} \quad [\text{using Lemma 2(b)}]$$

$$\to 0 \quad \text{as } A, n \to \infty.$$

(v) Three change points, $i_1 < i_2 \le m < i_3$: Let $\rho_{i_1 i_2 i_3}$ denote the partition with change points at i_1, i_2 and i_3:

$$\sum_{i_1} E_n\rho_{i_1 i_2 i_3} \le CE_n\rho_{i_2 i_3},$$

so it follows from (iv) that $E_n\{i_3 - i_2 > A\} \to 0$ as $A, n \to \infty$.

Since $\sum_{i_1, i_3 | i_1 < i_2 - A} E_n \rho_{i_1 i_2 i_3} \le C E_n \rho_{i_2} A^{-\alpha}$ and $\sum E_n \rho_{i_2}$ is bounded,

$$E_n\{i_1 < m - 2A\} \le E_n\{i_1 < i_2 - A\} + E_n\{i_2 < i_3 - A\} \to 0 \quad \text{as } A, n \to \infty.$$

We have proved that i_1 is asymptotically close to m in each of these five cases and all other cases are reducible to these, so the theorem is proved. □

We need to demonstrate that the powerful condition (2) used in the theorem is satisfied in interesting cases. Under suitable regularity conditions, the large sample behaviour of the densities $f_{0n}(X_{0n})$ is similar to the behaviour in sampling from a normal distribution. We will consider only this case in detail. The distribution P_n takes the observations $X_1, \ldots, X_m$ from $N(\theta_1, 1)$ and independently, the observations $X_{m+1}, \ldots, X_n$ from $N(\theta_2, 1)$.

Note that

$$\frac{(m + 1/\sigma_0^2)(n - m + 1/\sigma_0^2)}{n + 1/\sigma_0^2} \le \left(1 + \frac{1}{\sigma_0^2}\right)^2 \frac{m(n-m)}{n},$$

$$f_{0n}(X_{0n}) = (2\pi)^{-n/2}(1 + n\sigma_0^2)^{-1/2} \exp\left[-\frac{1}{2}\sum X_i^2 + \frac{n\hat{X}_{0n}^2/2}{1 + 1/(n\sigma_0^2)}\right],$$

$$\frac{f(\mathbf{X}|\rho)}{f(\mathbf{X}|\rho_m)} \le C\sigma_0^{-b} \exp\left[-\frac{m\hat{X}_{0m}^2/2}{1 + 1/(m\sigma_0^2)} - \frac{(n-m)\hat{X}_{mn}^2/2}{1 + 1/((n-m)\sigma_0^2)} + \sum_{ij \in \rho} \frac{(j-i)\hat{X}_{ij}^2/2}{1 + 1/((j-i)\sigma_0^2)}\right]\left(\frac{\sqrt{m(n-m)}}{\Pi_{ij \in \rho}\sqrt{j-i}}\right)$$

$$-\frac{m\hat{X}_{0m}^2}{1 + 1/(m\sigma_0^2)} - \frac{(n-m)\hat{X}_{mn}^2}{1 + 1/((n-m)\sigma_0^2)} + \sum_{ij \in \rho} \frac{(j-i)\hat{X}_{ij}^2}{1 + 1/((j-i)\sigma_0^2)}$$

$$\le (m - i_r)\hat{X}_{i_r m}^2 + \sum_{j \le m} (j-i)\hat{X}_{ij}^2 - m\hat{X}_{0m}^2 + (i_{r+1} - m)\hat{X}_{m i_{r+1}}^2$$

$$+ \sum_{i \ge m} (j-i)\hat{X}_{ij}^2 - (n-m)\hat{X}_{mn}^2 + (i_{r+1} - i_r)\hat{X}_{i_r i_{r+1}}^2$$

$$-(m - i_r)\hat{X}_{i_r m}^2 - (i_{r+1} - m)\hat{X}_{m i_{r+1}}^2$$

$$+ \sigma_0^2\left[-\hat{X}_{i_r i_{r+1}}^2 + \hat{X}_{i_r m}^2 + \hat{X}_{m i_{r+1}}^2\right].$$

The first six terms have their sum distributed as χ_{b-2}^2 under P_n. The remaining terms involve the change point. By the law of large numbers, for observations in a set Z_n where $P_n Z_n > 1 - \varepsilon_0$,

$$\sigma_0^2\left[-\hat{X}_{i_r i_{r+1}}^2 + \hat{X}_{i_r m}^2 + \hat{X}_{m i_{r+1}}^2\right] \le C \quad \text{for all } i_r, i_{r+1},$$

$$(i_{r+1} - i_r)\hat{X}_{i_r i_{r+1}}^2 - (m - i_r)\hat{X}_{i_r m}^2 - (i_{r+1} - m)\hat{X}_{m i_{r+1}}^2$$

$$\le C - \frac{1}{4}(\theta_2 - \theta_1)^2 \frac{(m - i_r)(i_{r+1} - m)}{i_{r+1} - i_r},$$

since $|\hat{X}_{mi_{r+1}} - \hat{X}_{i_r m}|$ is greater than $|\theta_2 - \theta_1|/2$ when both $m - i_r$ and $i_{r+1} - m$ are large.

Thus

$$P_n Z_n \left[\frac{f(\mathbf{X}|\rho)}{f(\mathbf{X}|\rho_m)}\right]^{\alpha} \leq C\sigma_0^{(1-b)\alpha}(1-\alpha)^{1/2(b-2)} \exp\left[C - \frac{1}{8}(\theta_2 - \theta_1)^2 \frac{(m - i_r)(i_{r+1} - m)}{i_{r+1} - i_r}\right] \times \left(\frac{m(n-m)}{\prod_{ij \in \rho}(j-i)}\right)^{1/2\alpha}.$$

Condition (2) holds with $\Delta = \exp(-(\theta_2 - \theta_1)^2/8)$ and $\delta = (1-\alpha)^{1/2}/\sigma_0^{\alpha}$. As long as the prior density is sufficiently diffuse, asymptotic consistency will hold for a single change point.

8. Selecting the number of change points. Yao (1988) uses a version of Schwarz's (1978) criterion to estimate the number of change points when the observations X_i, $i = 1, \ldots, n$, are independent normal $N(\mu_i, \sigma^2)$ with σ^2 unknown and with the sequence of means μ_i changing values at $R = b - 1$ change points. Yao treats the unknown change points as unknown parameters to be estimated, so that there are $2R + 2$ unknown parameters altogether, considering the unknown σ^2, the $R + 1 = b$ unknown means μ_i and the R unknown change points. The Schwarz criterion then selects that value of R that minimizes

$$\mathrm{SC}(R) = \tfrac{1}{2} n \log(\hat{\sigma}_R^2) + R \log(n),$$

where $\hat{\sigma}_R^2$ is the maximum likelihood estimate of σ^2 given R.

The Schwarz criterion is derived as an approximation to the probability of the data given a particular model, integrating out the unknown parameters. It would be interesting to know conditions on the distribution over partitions and of the distribution over parameters μ and σ, which would justify the above expression as an approximation to the probability of the data given R.

Yao justifies the use of the criterion in another way, showing that the value of R that maximizes the criterion converges to the true value of R with probability 1, provided that it is known that R is less than some given value R_0 and that the proportion of means in each block converges to some fixed fraction as $n \to \infty$.

In one way, Yao's result is stronger than statements of Theorem 1 and 2, which assert merely that if there is a single block, it will be discovered with probability 1, and if there are two blocks, *at least* two blocks will be discovered. In another way, our theorems are stronger, in that they assert that all the block boundaries will be close in posterior probability to the true block boundaries. It is also true that Yao's result applies only to the normal case [Yao (1988), page 188].

TABLE 1

For a sample of size 100, *number of times each number of change points was selected in* 100 *trials using Yao's method and the average posterior probabilities in the product partition method*

No change point					
Change points	0	1	2	3	4 +
Yao	89	6	5	0	0
Product partition	96	2	1	0	1
One change point, after observation 50, with $\mu_{51} - \mu_{50} = 1$					
Change points	0	1	2	3	4 +
Yao	01	78	20	1	0
Product partition	44	24	14	8	10
One change point, after observation 50, with $\mu_{51} - \mu_{50} = 2$					
Change points	0	1	2	3	4 +
Yao	0	90	9	1	0
Product partition	0	42	26	15	19

Nevertheless it may be instructive to compare Yao's method of selecting change points with the product partition method using the jump variable with density $g(i) = 4/(i(i + 1)(i + 2))$. In Table 1, we compare the number of times, in 100 trials, that Yao's method selected various numbers of change points, with the average posterior probabilities for those numbers of change points under the product partition method; we selected $\sigma_0^2 = 16$, $\mu_0 = \bar{X}$ and took σ^2 to be the maximum likelihood estimate. The sample size was 100.

It can be seen that Yao's method is better at identifying the number of groups accurately. As the size of the shift increases, Yao will more and more surely select only two groups, whereas the product method using these prior cohesions will always allow for the possibility that there may be more groups. Perhaps this defect can be cured by specifying cohesions that inhibit the formation of small groups; for example, in Yao's selection criterion the number of groups is bounded and all groups are required to be order n in size. But in practical problems, we are often interested in small groups with sufficiently large deviations, such as outliers, so an outright prohibition of small groups should be avoided. In addition, errors in estimating means at each point are serious when the group endpoints are inaccurately estimated and these errors are less likely if the number of groups is overestimated rather than underestimated or exactly estimated; more extensive simulation study has shown that the mean square errors in estimating means are notably higher for the Schwarz selection method than appropriate product partition methods, even though the product partition method overestimates the number of groups.

REFERENCES

BARNARD, G. A. (1959). Control charts and stochastic processes (with discussion). *J. Roy. Statist. Soc. Ser. B* **21** 239–271.

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

DEMPSTER, A. P., LAIRD, N. M. and RUBIN D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DUNCAN, A. J. (1956). The economic design of $\bar{X}$ charts used to maintain current control of a process. *J. Amer. Statist. Assoc.* **51** 228–242.

HARTIGAN, J. A. (1983). *Bayes Theory*. Springer, New York.

HARTIGAN, J. A. (1990). Partition models. *Comm. Statist. Theory Methods* **19** 2745–2756.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Harvard Univ. Press.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

YAO, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *Ann. Statist.* **12** 1434–1447.

YAO, Y. C. (1988). Estimating the number of change-points by Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–187.

DEPARTMENT OF STATISTICS
UNIVERSITY COLLEGE
CORK
IRELAND

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520