Ann. Inst. Statist. Math.
Vol. 43, No. 1, 77–93 (1991)

BAYESIAN DETECTION OF STRUCTURAL CHANGES

NOBUHISA KASHIWAGI

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan

(Received November 16, 1988; revised February 21, 1990)

Abstract. A Bayesian solution is given to the problem of making inferences about an unknown number of structural changes in a sequence of observations. Inferences are based on the posterior distribution of the number of change points and on the posterior probabilities of possible change points. Detailed analyses are given for binomial data and some regression problems, and numerical illustrations are provided. In addition, an approximation procedure to compute the posterior probabilities is presented.

Key words and phrases: Bayesian inference, change point, predictive log likelihood, Lindisfarne scribes problem, regression.

1. Introduction

Let $y_1, \ldots, y_T$ be a sequence of observations taken at equally spaced intervals. A sequence of random variables $Y_1, \ldots, Y_T$ is said to have n change points at $j(1), \ldots, j(n)$ $(1 \leq j(1) < \cdots < j(n) < T)$ if the density of $\boldsymbol{y} = (y_1, \ldots, y_T)'$ has the form

$$p(\boldsymbol{y} \mid J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_n) = \prod_{i=0}^{n} p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta}_i) \tag{1.1}$$

where $J_{j(i)}$ is the event that a sequence has a change point at $j(i)$; N is the number of change points; $\boldsymbol{y}_i = (y_{j(i)+1}, \cdots, y_{j(i+1)})'$, $j(0) = 0$, $j(n+1) = T$; $p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta}_i)$ is the density of $\boldsymbol{y}_i$ with a parameter $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_{i'}$ $i \neq i'$. In this paper, we consider the problem of making inferences about change points under the conditions that the places of change points, the number of change points and the values of $\boldsymbol{\theta}_i$'s are unknown.

Since Page (1954), the change point problem has been considered by many authors from various viewpoints. For example, changes in a sequence of random variables have been considered by Bhattacharya and Johnson (1968), Pettitt (1979), Schechtman and Wolfe (1985), Lombard (1987) and Carlstein (1988) from the nonparametric viewpoint; by Hinkley (1970) and James *et al.* (1987) from the maximum likelihood viewpoint; and by Chernoff and Zacks (1964) and Smith

(1975) from the Bayesian viewpoint. Changes in the regression case have been considered by Quandt (1958, 1960) and Hinkley (1969, 1971) from the maximum likelihood viewpoint; and by Bacon and Watts (1971), Booth and Smith (1982) and Tsurumi *et al.* (1986) from the Bayesian viewpoint. Changes in the time-series case have been considered by Box and Tiao (1965), Harrison and Stevens (1976) and Kitagawa (1987) from the Bayesian viewpoint. (For other references, see Poirier (1976), Zacks (1983) and Broemeling and Tsurumi (1987)).

However, most studies have been concerned with the detection of a single change or the detection of more than one change using a stepwise procedure; as a result, few studies are available on the problem of detecting more than one change without using a stepwise procedure. Smith (1980) is one of those few studies. He analyzed the Lindisfarne scribes problem and gave posterior probabilities of up to two changes.

In this paper, we deal with the problem of detecting more than one change point without using a stepwise procedure from the Bayesian viewpoint. For this, we propose a method to evaluate the posterior distribution of N and the posterior probability of each J_t unconditionally, using the predictive log likelihood proposed by Kitagawa and Akaike (1982). We also present an approximation procedure for decreasing the amount of computation.

In Section 2, a Bayesian formulation of the problem is presented. In Section 3, the detailed analysis is given for binomial data and the Lindisfarne scribes problem is analyzed. In Section 4, an approximation procedure is presented. In Section 5, changes in the regression case are studied for two specific models, the simple regression model and the discrete spline; numerical illustrations are also provided.

2. A Bayesian formulation

In this section, we derive the posterior distribution of N and the posterior probability of J_t.

When the sequence is assumed to have n change points at $j(1), \ldots, j(n)$, the density of $\boldsymbol{y}$ is given by (1.1). Assuming a prior density $\omega(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$; where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_n)$, the integrated likelihood of $\{J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n\}$ is written as

$$p(\boldsymbol{y} \mid J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n) = \int \cdots \int p(\boldsymbol{y} \mid J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n, \boldsymbol{\theta}) \omega(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

By Bayes' theorem, the posterior probability of $J_{j(1)} \cap \cdots \cap J_{j(n)}$ given $\boldsymbol{y}$ and n is provided as

$$p(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid \boldsymbol{y}, N = n) = \frac{p(\boldsymbol{y} \mid J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n) \omega(J_{j(1)} \cap \cdots \cap j_{j(n)} \mid N = n)}{p(\boldsymbol{y} \mid N = n)}$$

where $\omega(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid N = n)$ is a prior probability of $J_{j(1)} \cap \cdots \cap J_{j(n)}$ given n and

$$p(\boldsymbol{y} \mid N = n)$$
$$= \sum_{\Omega_n} p(\boldsymbol{y} \mid J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n)\omega(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid N = n)$$
$$\Omega_n = \{(j(1), \ldots, j(n)) \mid 1 \leq j(1) < \cdots < j(n) < T\}.$$

Taking the sum of $p(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid \boldsymbol{y}, N = n)$'s which involve J_t, the posterior probability of J_t given $\boldsymbol{y}$ and n is obtained as

$$p(J_t \mid \boldsymbol{y}, N = n) = \sum_{\Omega_{n,t}} p(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid \boldsymbol{y}, N = n)$$
$$\Omega_{n,t} = \{(j(1), \ldots, j(n)) \mid \exists k \quad \text{such that } j(k) = t$$
$$1 \leq k \leq n, \quad 1 \leq j(1) < \cdots < j(n) < T\}.$$

On the other hand, the posterior probability of $N = n$ given $\boldsymbol{y}$ is provided by Bayes' theorem as

$$p(N = n \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid N = n)\omega(N = n)}{p(\boldsymbol{y})}$$

where $\omega(N = n)$ is a prior probability of $N = n$ and

$$p(\boldsymbol{y}) = \sum_{n=0}^{T-1} p(\boldsymbol{y} \mid N = n)\omega(N = n).$$

The posterior probability of J_t given $\boldsymbol{y}$ is obtained as

$$p(J_t \mid \boldsymbol{y}) = \sum_{n=1}^{T-1} p(J_t \mid \boldsymbol{y}, N = n)p(N = n \mid \boldsymbol{y}).$$

The necessary ingredients to evaluate the posterior probabilities in the above formulation are $p(\boldsymbol{y} \mid J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n)$, $\omega(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid N = n)$ and $\omega(N = n)$. In this paper, as $\omega(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid N = n)$ and $\omega(N = n)$ we assume the following prior probabilities used in Smith (1980) and Kitagawa and Akaike (1982).

$$\begin{cases} \omega(J_{j(1)} \cap \cdots \cap J_{j(n)} \mid N = n) = \dfrac{1}{{}_{T-1}C_n} & 1 \leq n < T \\ \omega(N = n) = \dfrac{1}{T} & 0 \leq n < T. \end{cases}$$

The remaining ingredient, the likelihood of $J = \{J_{j(1)} \cap \cdots \cap J_{j(n)}, N = n\}$ is provided concretely for some models in Sections 3, 5.1 and 5.2. In these sections, following Kitagawa and Akaike (1982) and according to the entropy maximization principle (Akaike (1977)), the model $p(\boldsymbol{y} \mid J, \boldsymbol{\theta})$ is specified by the maximum

likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Then the goodness of the model is evaluated by its expected log likelihood $E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}})$; where E_Z denotes the expectation under the assumed distribution of $\boldsymbol{Z}$, $p(\boldsymbol{z} \mid J, \boldsymbol{\theta})$. As an estimate of the expected log likelihood, the predictive log likelihood is used, which is defined by

$$(2.1)\quad \log p^{\text{pred}}(\boldsymbol{y} \mid J) = \log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}) - E_Y[\log p(\boldsymbol{Y} \mid J, \hat{\boldsymbol{\theta}}) - E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}})]$$

where E_Y denotes the expectation under the assumed distribution of data. Using (2.1), we define the likelihood of J.

3. Lindisfarne scribes problem

The Lindisfarne scribes problem is one of the well-known examples of the change point problem. The aim in this problem is to make inferences about changes of scribe using the data on the number of occurrences of present indicative 3rd singular endings s and δ in each section of Lindisfarne. Table 1 shows the data taken from Smith (1980). These data have been analyzed by Smith (1980), Silvey (1958), Pettitt (1979) and Carlstein (1988). The latter three authors drew the conclusion using some test statistics that the change occurred after the 5th section. Smith (1980) evaluated the posterior probabilities of up to two changes and concluded that the change occurred after the 4th section and again after the 5th section. In this section, we apply our method to the data of Lindisfarne and compare our results with theirs.

Table 1. Number of occurrences of present indicative 3rd singular endings s and δ for different sections of Lindisfarne.

Section	s	δ	Total
1	12	9	21
2	26	10	36
3	31	13	44
4	24	6	30
5	28	24	52
6	34	11	45
7	39	9	48
8	46	11	57
9	41	7	48
10	19	3	22
11	17	3	20
12	17	4	21
13	16	4	20

Let m_t, y_t be the numbers of occurrences of present indicative 3rd singular endings and δ-forms at the t-th section ($t = 1, \ldots, 13$), respectively. Similarly to

Smith (1980), we assume the binomial distribution with parameters m_t, θ_i for y_t $(t = j(i)+1, \ldots, j(i+1),\ i = 0, \ldots, n)$. Then the integrated likelihood of J is written as

$$p(\boldsymbol{y} \mid J) = \int \cdots \int \prod_{i=0}^{n} \prod_{t=j(i)+1}^{j(i+1)} {}_{m_t}C_{y_t} \theta_i^{y_t} (1-\theta_i)^{m_t - y_t} \omega(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_n)$.

As $\omega(\boldsymbol{\theta})$, Smith (1980) assumed a conjugate prior distribution. This is one of several possible selections. But in this paper, as mentioned in the previous section, we specify the model by the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Then we derive the predictive log likelihood (2.1) to define the likelihood of J.

From the assumption for y_t, the maximum likelihood estimate of θ_i is obtained as $\hat{\theta}_i = \sum_t y_t / \sum_t m_t$ $(i = 0, \ldots, n)$, and the maximum log likelihood is obtained as

$$\log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}) = \sum_{t=1}^{13} \log {}_{m_t}C_{y_t} + \sum_{i=0}^{n} \sum_{t=j(i)+1}^{j(i+1)} \{y_t \log \hat{\theta}_i + (m_t - y_t) \log(1-\hat{\theta}_i)\}.$$

It may be a possible selection to use the maximum likelihood as an estimate of $p(\boldsymbol{y} \mid J)$. However, the maximum log likelihood has the bias

$$\begin{aligned} &\log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}) - E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}}) \\ &\quad = \sum_{t=1}^{13} (\log {}_{m_t}C_{y_t} - E_Z \log {}_{m_t}C_{Z_t}) + \sum_{i=0}^{n} f_i (\hat{\theta}_i - \theta_i) \log \frac{\hat{\theta}_i}{(1-\hat{\theta}_i)} \end{aligned}$$

where $f_i = \sum_{t=j(i)+1}^{j(i+1)} m_t$ $(i = 0, \ldots, n)$. This bias increases in average as $\dim(\boldsymbol{\theta})$ becomes large, which means that the use of the maximum likelihood causes an overestimation of the number of change points. To prevent such an overestimation, it is necessary to correct the bias. However, since the true parameter $\boldsymbol{\theta}$ is unknown, the present form of the bias is useless. Consequently, we employ the predictive log likelihood (2.1).

The expectation of the bias can be written as

$$\begin{aligned} &E_Y[\log p(\boldsymbol{Y} \mid J, \hat{\boldsymbol{\theta}}) - E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}})] \\ &\quad = \sum_{i=0}^{n} \left\{ 1 + \frac{\theta_i^2 - \theta_i + \frac{1}{2}}{f_i \theta_i (1-\theta_i)} + \frac{\theta_i^4 - 2\theta_i^3 + 4\theta_i^2 - 3\theta_i + \frac{5}{6}}{f_i^2 \theta_i^2 (1-\theta_i)^2} + O(f_i^{-3}) \right\}. \end{aligned}$$

Using this form, we define the predictive log likelihood in the current problem as

$$\begin{aligned} &\log p^{\mathrm{pred}}(\boldsymbol{y} \mid J) \\ &\quad = \log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}) - \sum_{i=0}^{n} \left\{ 1 + \frac{\hat{\theta}_i^2 - \hat{\theta}_i + \frac{1}{2}}{f_i \hat{\theta}_i (1-\hat{\theta}_i)} + \frac{\hat{\theta}_i^4 - 2\hat{\theta}_i^3 + 4\hat{\theta}_i^2 - 3\hat{\theta}_i + \frac{5}{6}}{f_i^2 \hat{\theta}_i^2 (1-\hat{\theta}_i)^2} \right\}. \end{aligned}$$

Table 2. Posterior distribution of N and Smith's results

n	$p(N = n \mid \boldsymbol{y})$
0	0.003
1	0.185
2	0.210
3	0.194
4	0.155
5	0.109
6	0.068
7	0.038
8	0.020
9	0.010
10	0.004
11	0.002
12	0.001
Mean	3.4
Mode	2
Median	3

	Smith's
n	$p(N = n \mid \boldsymbol{y})$
0	0.000
1	0.069
2	0.931

Table 3. Posterior probabilities of J_t's.

t	$p(J_t \mid \boldsymbol{y})$
1	0.265
2	0.176
3	0.215
4	0.544
5	0.744
6	0.382
7	0.205
8	0.210
9	0.158
10	0.151
11	0.158
12	0.146

As the estimate of $p(\boldsymbol{y} \mid J)$, we use $\exp\{\log p^{\text{pred}}(\boldsymbol{y} \mid J)\}$.

Now we apply our method to the data of Lindisfarne. Table 2 presents the estimate of each $p(N = n \mid \boldsymbol{y})$ as well as the posterior mean, mode and median of N and Smith's results. Table 3 presents the estimate of each $p(J_t \mid \boldsymbol{y})$. It is difficult to precisely compare our results with those of Smith because he has not presented the posterior probabilities of more than two changes; nevertheless, there seem to be some differences between them. While the posterior probability of two changes is quite dominant in Smith's results, it is not so dominant in our results. This difference may be caused by the difference between the assumed distributions for $\boldsymbol{\theta}$ and by the different policy for the bias correction. However, in spite of this difference between both results, we can agree with Smith's conclusion. Actually, if we take the posterior mode of N, the conclusion that there are two changes is obtained. From Table 3, it is seen that the top two $p(J_t \mid \boldsymbol{y})$'s are obtained at the 4th and 5th sections.

4. An approximation procedure

We call the evaluation of the posterior probabilities by the method mentioned in Section 2 *the full computation*. In the Lindisfarne scribes problem, *the full computation* was feasible. However, the number of estimations of $p(\boldsymbol{y} \mid J)$'s in *the full computation*, which is given by $\sum_{n=0}^{T-1} {}_{T-1}C_n$, increases exponentially with the size of the sequence and quickly *the full computation* becomes infeasible. In this section, we present an approximation procedure which enables us to evaluate $p(J_t \mid \boldsymbol{y})$'s even when *the full computation* is infeasible.

The flow of the approximation is as follows:

0. Calculate $p(\boldsymbol{y} \mid N = 0)$, and set $n \leftarrow 1$.

1. Calculate $p(J_t \mid \boldsymbol{y}, N = n)$ $(1 \leq t < T)$ by the method in Section 2.

2. Let m be the number of repetitions of Step 1. If $n < m$ then set $n \leftarrow n+1$ and return to Step 1.

3. Determine whether n is sufficiently large to terminate. If so, then go to Step 7. If not, then set $n \leftarrow n + 1$.

4. Let α be a small value. Make the index set $I_{n,\alpha} = \{i \mid p(J_i \mid \boldsymbol{y}, N = n-1) \leq \alpha \;\; 1 \leq i < T\}$ and calculate $g(n, t) \equiv \sum_{\Omega_{n,t}} p(\boldsymbol{y} \mid J)$ $(1 \leq t < T)$ under the following assumption:

$$g(n, t) = \begin{cases} \dfrac{{}_{T-2}C_{n-1}\, g(n-1, t)}{{}_{T-2}C_{n-2}} & t \in I_{n,\alpha} \\ \displaystyle\sum_{\Omega^{\alpha}_{n,t}} p(\boldsymbol{y} \mid J) + \sum_{k \in I_{n,\alpha}} \frac{p(J_t \mid \boldsymbol{y}, N = n-1)\, g(n, k)}{1 - \dfrac{p(J_k \mid \boldsymbol{y}, N = n-1)}{n-1}} & t \notin I_{n,\alpha} \end{cases}$$

$$\Omega^{\alpha}_{n,t} = \{(j(1), \ldots, j(n)) \mid (j(1), \ldots, j(n)) \in \Omega_{n,t},\; j(i) \notin I_{n,\alpha} \; 1 \leq i \leq n\}.$$

5. Using $g(n, t)$'s and the relations

$$p(J_t \mid \boldsymbol{y}, N = n) = \frac{g(n, t)}{{}_{T-1}C_n p(\boldsymbol{y} \mid N = n)}$$

$$p(\boldsymbol{y} \mid N = n) = \frac{1}{{}_{T-1}C_n n} \sum_{t=1}^{T-1} g(n, t),$$

calculate $p(J_t \mid \boldsymbol{y}, N = n)$ $(1 \leq t < T)$.

6. Return to Step 3.

7. Calculate $p(J_t \mid \boldsymbol{y})$ $(1 \leq t < T)$ assuming the prior

$$\omega(N = k) = \begin{cases} \dfrac{1}{n+1} & k \leq n \\ 0 & k > n. \end{cases}$$

The number of estimations of $p(\boldsymbol{y} \mid J)$'s is decreased in Step 4 by approximating $g(n, t)$'s. The approximation of $g(n, t)$ is introduced as follows. Consider the case where J_k is assumed to be an unimportant event, that is $I_{n,\alpha} = \{k\}$. In this case, it may be reasonable to consider assigning approximate values to the predictive likelihoods of J's which involve J_k in order to decrease the amount of computation. To obtain such approximate values, we set the following two assumptions. The first assumption is that the mean of the predictive likelihoods of J's which involve J_k when $N = n$ is equivalent to the mean of those when $N = n - 1$. By this assumption, we have

$$g(n, k) = \frac{{}_{T-2}C_{n-1} g(n-1, k)}{{}_{T-2}C_{n-2}}.$$

On the other hand, $g(n, t)$ $(t \neq k)$ can be written as

$$g(n, t) = g_1(n, t) + g_2(n, t)$$

$$g_1(n, t) = \sum_{\Omega^{\alpha}_{n,t}} p(\boldsymbol{y} \mid J), \qquad g_2(n, t) = \sum_{\Omega_{n,t} - \Omega^{\alpha}_{n,t}} p(\boldsymbol{y} \mid J).$$

We evaluate $g_1(n, t)$ by the method mentioned in Section 2. However, since $g_2(n, t)$ is the sum of the predictive likelihoods of J's which involve J_k, we assign an approximate value to it. An approximate value can be obtained using the relation

$$\sum_{t \neq k} g_2(n, t) = (n-1) g(n, k).$$

This relation suggests that we may distribute $(n-1)g(n, k)$ into $g_2(n, t)$'s $(t \neq k)$. Using the posterior probabilities when $N = n - 1$, we set the second assumption

$$g_2(n, t) = \frac{p(J_t \mid \boldsymbol{y}, N = n-1)/(n-1)}{1 - \{p(J_k \mid \boldsymbol{y}, N = n-1)/(n-1)\}} (n-1) g(n, k) \qquad t \neq k.$$

The approximation of $g(n, t)$ has been obtained.

The above two assumptions may be ad hoc. However, a close approximation increases the amount of computation. We consider that they are acceptable ones in practical application.

In the approximation procedure, there are some arbitrary constants, m and α. A basic strategy as to their choice is to chose the largest m and smallest α, as large and small, respectively, as the computer may permit. By some experiments, we have found that: (i) When m is greater than the mode of N, the possibility to miss change points is very small. (ii) When α is less than a certain value, as the number of elements of $I_{n,\alpha}$ is less than about $T-n-6$, relatively good approximate values are obtained.

5. Detection of changes by regression models

In this section, we give the estimate of $p(\boldsymbol{y} \mid J)$ for two regression models, the simple regression model and the discrete spline. In addition, we show an example of the application of our method using the discrete spline.

5.1 *The simple regression model*

Assume the simple regression model

$$y_t = \alpha_i + \beta_i t + \epsilon_t \qquad \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2) \qquad j(i)+1 \leq t \leq j(i+1)$$

for $\boldsymbol{y}_i$. Then the density of $\boldsymbol{y}_i$ can be written as

$$p_S(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \sigma^2) = (2\pi)^{-\kappa_i/2}\sigma^{-\kappa_i} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - A_i\boldsymbol{\mu}_i)'(\boldsymbol{y}_i - A_i\boldsymbol{\mu}_i)\right\} \tag{5.1}$$

where $\boldsymbol{\mu}_i = (\alpha_i, \beta_i)'$, $\kappa_i = \dim(\boldsymbol{y}_i)$ and

$$A_i = \begin{bmatrix} 1 & j(i)+1 \\ 1 & j(i)+2 \\ \vdots & \vdots \\ 1 & j(i+1) \end{bmatrix}.$$

Further, since the simple regression model is inapplicable to $\boldsymbol{y}_i$ when $\kappa_i = 1$, we assume the following outlier model for such $\boldsymbol{y}_i$.

$$p_N(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{y_{j(i)+1} - \alpha_i}{\sigma}\right) \tag{5.2}$$

where $\boldsymbol{\mu}_i = (\alpha_i)$ and ϕ denotes the standard normal probability density function. Using (5.1) and (5.2), the density of $\boldsymbol{y}$ can be written as

$$p(\boldsymbol{y} \mid J, \boldsymbol{\theta}) = \prod_{i \in I_1} p_N(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \sigma^2) \prod_{i \in I_G} p_S(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \sigma^2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}_0', \ldots, \boldsymbol{\mu}_n', \sigma^2)$, $I_1 = \{i \mid \kappa_i = 1 \quad 0 \leq i \leq n\}$ and $I_G = \{i \mid \kappa_i \geq 2 \quad 0 \leq i \leq n\}$. Therefore the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained as

$$\hat{\alpha}_i = y_{j(i)+1} \qquad \text{for} \quad i \in I_1$$

$$\hat{\boldsymbol{\mu}}_i = (A_i'A_i)^{-1}A_i'\boldsymbol{y}_i \qquad \text{for} \quad i \in I_G$$

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{i \in I_G}(\boldsymbol{y}_i - A_i\hat{\boldsymbol{\mu}}_i)'(\boldsymbol{y}_i - A_i\hat{\boldsymbol{\mu}}_i)$$

and the maximum log likelihood is obtained as

$$\log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}) = -\frac{T}{2}\log 2\pi\hat{\sigma}^2 - \frac{T}{2}.$$

The bias of the maximum log likelihood is given by

$$\log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}) - E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}})$$
$$= \frac{1}{2\hat{\sigma}^2}\left\{T\sigma^2 + \sum_{i \in I_1}(\alpha_i - \hat{\alpha}_i)^2 + \sum_{i \in I_G}(\mu_i - \hat{\mu}_i)'A_i'A_i(\mu_i - \hat{\mu}_i)\right\} - \frac{T}{2}.$$

Since

$$\frac{(\alpha_i - \hat{\alpha}_i)^2}{\sigma^2} \sim \chi^2_{(1)} \qquad \text{for} \quad i \in I_1$$

$$\frac{(\mu_i - \hat{\mu}_i)'A_i'A_i(\mu_i - \hat{\mu}_i)}{\sigma^2} \sim \chi^2_{(2)} \qquad \text{for} \quad i \in I_G$$

$$\frac{T\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{\left(\sum_{i \in I_G} k_i - 2\#I_G\right)},$$

we have

$$E_Y[\log p(\boldsymbol{Y} \mid J, \hat{\boldsymbol{\theta}}) - E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}})] = \frac{T(T + \#I_1 + 2\#I_G)}{2\left(\sum_{i \in I_G}\kappa_i - 2\#I_G - 2\right)} - \frac{T}{2}$$

where $\chi^2_{(k)}$ denotes the χ^2-distribution of order k and $\#I_*$ the number of elements included in the set I_*. Therefore the predictive log likelihood is obtained as

$$\log p^{\text{pred}}(\boldsymbol{y} \mid J) = -\frac{T}{2}\log 2\pi\hat{\sigma}^2 - \frac{T(T + \#I_1 + 2\#I_G)}{2\left(\sum_{i \in I_G}\kappa_i - 2\#I_G - 2\right)}.$$

The estimate of the likelihood $p(\boldsymbol{y} \mid J)$ is obtained as $\exp\left\{\log p^{\text{pred}}(\boldsymbol{y} \mid J)\right\}$.

5.2 *The discrete spline*

Harrison and Stevens (1976) presented three examples of sequences including a single change, which are shown in Fig. 1. Although they generated these sequences by the linear growth model, it is possible to represent them by the model mentioned in the previous section. For example, the outlier case can be represented by applying model (5.1) to the data at $1 \leq t \leq 4$ and $6 \leq t \leq 10$ and applying model (5.2) to the data at $t = 5$. However, if the data at $1 \leq t \leq 4$ and $6 \leq t \leq 10$ are on a curve instead of a straight line, the model mentioned in the previous section becomes inappropriate. For such a case, a state-space model for regression, called the discrete spline by Tanabe and Tanaka (1983), is useful.

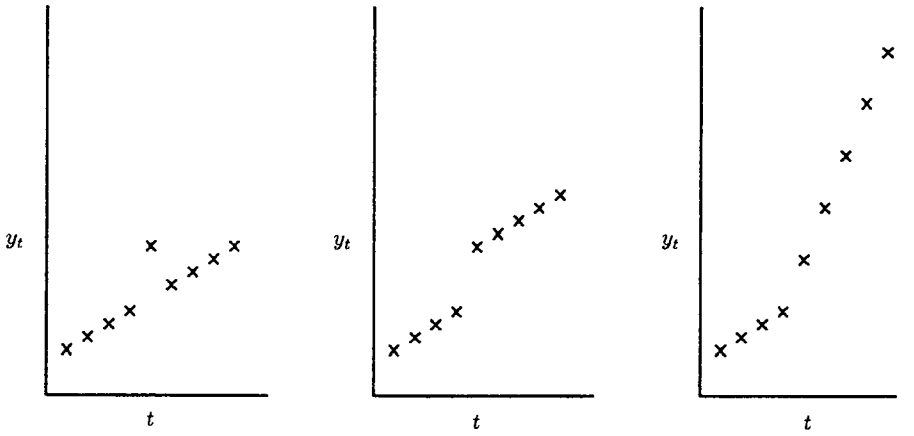

Fig. 1. Examples of sequences including a single change.

The discrete spline for $\boldsymbol{y}_i$ is defined by

$$(5.3)\quad y_t = x_t + \epsilon_t \qquad \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2) \qquad j(i)+1 \leq t \leq j(i+1)$$

$$(5.4)\quad x_t = 2x_{t-1} - x_{t-2} + \zeta_t \qquad \zeta_t \sim \text{i.i.d. } N\left(0, \frac{\sigma^2}{\lambda^2}\right) \qquad j(i)+3 \leq t \leq j(i+1)$$

where x_t denotes the trend. From (5.3), the density of $\boldsymbol{y}_i$ is given by

$$p(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\nu}_i, \sigma^2) = (2\pi)^{-\kappa_i/2}\sigma^{-\kappa_i} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{x}_i)'(\boldsymbol{y}_i - \boldsymbol{x}_i)\right\}$$

where $\boldsymbol{\mu}_i = (x_{j(i)+1}, x_{j(i)+2})'$, $\boldsymbol{\nu}_i = (x_{j(i)+3}, \ldots, x_{j(i+1)})'$ and $\boldsymbol{x}_i = (\boldsymbol{\mu}_i', \boldsymbol{\nu}_i')'$. From (5.4), the prior density of $\boldsymbol{\nu}_i$ is given by

$$\omega(\boldsymbol{\nu}_i \mid \boldsymbol{\mu}_i, \sigma^2, \lambda) = (2\pi)^{-(\kappa_i-2)/2}\left(\frac{\lambda}{\sigma}\right)^{\kappa_i-2} \exp\left\{-\frac{\lambda^2}{2\sigma^2}(D\boldsymbol{x}_i)'D\boldsymbol{x}_i\right\}$$

$$D = \begin{bmatrix} 1 & -2 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & -2 & 1 \end{bmatrix}.$$

Therefore the density of $\boldsymbol{y}_i$ corresponding to the discrete spline is given by

$$
\begin{aligned}
(5.5)\quad & p_D(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \sigma^2, \lambda^2) \\
&= \int p(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\nu}_i, \sigma^2)\omega(\boldsymbol{\nu}_i \mid \boldsymbol{\mu}_i, \sigma^2, \lambda)d\boldsymbol{\nu}_i \\
&= (2\pi)^{-\kappa_i/2}\sigma^{-\kappa_i} \mid V_i \mid^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - A_i\boldsymbol{\mu}_i)'V_i^{-1}(\boldsymbol{y}_i - A_i\boldsymbol{\mu}_i)\right\}
\end{aligned}
$$

where $V_i = I_{\kappa_i} + (1/\lambda^2)B_iB_i'$, I_{κ_i} denotes the identity matrix of rank κ_i and

$$
A_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -2 \\ \vdots & \vdots \\ \kappa_i - 2 & 1 - \kappa_i \end{bmatrix} \qquad B_i = \begin{bmatrix} 0 & 0 & & 0 \\ 0 & 0 & & \\ 1 & 0 & & \\ 2 & 1 & & \\ \vdots & \vdots & \vdots & \\ \kappa_i - 2 & \kappa_i - 3 & \cdots & 1 \end{bmatrix}.
$$

On the other hand, since the discrete spline is inapplicable to $\boldsymbol{y}_i$ when $\kappa_i \leq 2$, we assume the following model for such $\boldsymbol{y}_i$.

$$
(5.6)\qquad p_N(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \sigma^2) = \prod_{k=j(i)+1}^{j(i+1)} \frac{1}{\sigma}\phi\left(\frac{y_k - \alpha_i}{\sigma}\right)
$$

where $\boldsymbol{\mu}_i = (\alpha_i)$. Using (5.5) and (5.6), the density of $\boldsymbol{y}$ can be written as

$$
p(\boldsymbol{y} \mid J, \boldsymbol{\theta}, \lambda) = \prod_{i \in I_1 \cup I_2} p_N(\boldsymbol{y}_i \mid \mu_i, \sigma^2) \prod_{i \in I_G} p_D(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \sigma^2, \lambda)
$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}_0', \ldots, \boldsymbol{\mu}_n', \sigma^2)$, $I_1 = \{i \mid \kappa_i = 1 \quad 0 \leq i \leq n\}$, $I_2 = \{i \mid \kappa_i = 2 \quad 0 \leq i \leq n\}$ and $I_G = \{i \mid \kappa_i \geq 3 \quad 0 \leq i \leq n\}$. In this model, the maximum likelihood estimate of λ is hard to obtain analytically. Consequently, we first assume that λ is fixed. Then we obtain the conditional maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ as

$$
\begin{aligned}
&\hat{\alpha}_i = y_{j(i)+1} && \text{for} \quad i \in I_1 \\
&\hat{\alpha}_i = \frac{y_{j(i)+1} + y_{j(i)+2}}{2} && \text{for} \quad i \in I_2 \\
&\hat{\boldsymbol{\mu}}_i = (A_i'V_i^{-1}A_i)^{-1}A_i'V_i^{-1}\boldsymbol{y}_i && \text{for} \quad i \in I_G \\
&\hat{\sigma}^2 = \frac{1}{T}\left\{\sum_{i \in I_2}\sum_{k=j(i)+1}^{j(i+1)} (y_k - \hat{\alpha}_i)^2 + \sum_{i \in I_G}(\boldsymbol{y}_i - A_i\hat{\boldsymbol{\mu}}_i)'V_i^{-1}(\boldsymbol{y}_i - A_i\hat{\boldsymbol{\mu}}_i)\right\}
\end{aligned}
$$

and the conditional maximum log likelihood as

$$
\log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}, \lambda) = -\frac{T}{2}\log 2\pi\hat{\sigma}^2 - \frac{1}{2}\sum_{i \in I_G}\log \mid V_i \mid - \frac{T}{2}.
$$

The bias of the conditional maximum log likelihood is given by

$$\log p(\boldsymbol{y} \mid J, \hat{\boldsymbol{\theta}}, \lambda) - E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}}, \lambda)$$
$$= \frac{1}{2\hat{\sigma}^2}\left\{T\sigma^2 + \sum_{i\in I_1}(\alpha_i - \hat{\alpha}_i)^2 + \sum_{i\in I_2} 2(\alpha_i - \hat{\alpha}_i)^2 + \sum_{i\in I_G}(\mu_i - \hat{\mu}_i)' A_i' V_i^{-1} A_i (\mu_i - \hat{\mu}_i)\right\} - \frac{T}{2}.$$

Since

$$\frac{(\alpha_i - \hat{\alpha}_i)^2}{\sigma^2} \sim \chi^2_{(1)} \qquad \text{for} \quad i \in I_1$$

$$\frac{2(\alpha_i - \hat{\alpha}_i)^2}{\sigma^2} \sim \chi^2_{(1)} \qquad \text{for} \quad i \in I_2$$

$$\frac{(\mu_i - \hat{\mu}_i)' A_i' V_i^{-1} A_i (\mu_i - \hat{\mu}_i)}{\sigma^2} \sim \chi^2_{(2)} \qquad \text{for} \quad i \in I_G$$

$$\frac{T\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{\left(\#I_2 + \sum_{i\in I_G}\kappa_i - 2\#I_G\right)},$$

we have

$$E_Y[\log p(\boldsymbol{Y} \mid J, \hat{\boldsymbol{\theta}}, \lambda) - E_Z \log p(\boldsymbol{Z} \mid J, \hat{\boldsymbol{\theta}}, \lambda)]$$
$$= \frac{T(T + \#I_1 + \#I_2 + 2\#I_G)}{2\left(\#I_2 + \sum_{i\in I_G}\kappa_i - 2\#I_G - 2\right)} - \frac{T}{2}.$$

Therefore the conditional predictive log likelihood is obtained as

$$\log p^{\text{pred}}(\boldsymbol{y} \mid J, \lambda)$$
$$= -\frac{T}{2}\log 2\pi\hat{\sigma}^2 - \frac{1}{2}\sum_{i\in I_G}\log \mid V_i \mid - \frac{T(T + \#I_1 + \#I_2 + 2\#I_G)}{2\left(\#I_2 + \sum_{i\in I_G}\kappa_i - 2\#I_G - 2\right)}.$$

Assuming a prior density $\omega(\lambda)$ for λ, we obtain the estimate of $p(\boldsymbol{y} \mid J)$ as

$$\int \exp\left\{\log p^{\text{pred}}(\boldsymbol{y} \mid J, \lambda)\right\}\omega(\lambda)d\lambda.$$

In the actual numerical computation, we usually assume for λ a prior distribution having uniform probabilities on finite discrete points, for example, $\omega(\lambda) = 1/8$ ($\lambda =$ 1, 2, 4, 8, 16, 32, 64, 128).

5.3 *An example of application*

In this section, we apply our method using the discrete spline to the data of opinion polls on the proportion of voters who support the Japan Liberal Democratic Party collected by Chuochosa-sha, a Japanese institute conducting sample

surveys every month from December 1978 to November 1982. Figure 2 shows the data plotted against time. In this example, since *the full computation* is infeasible, we use the approximation procedure under the following conditions: (I) m is set as $m = 4$. (II) α is set as $\alpha = 0.004n$. (III) The procedure is terminated at $n = 9$. The results are given in Tables 4 and 5.

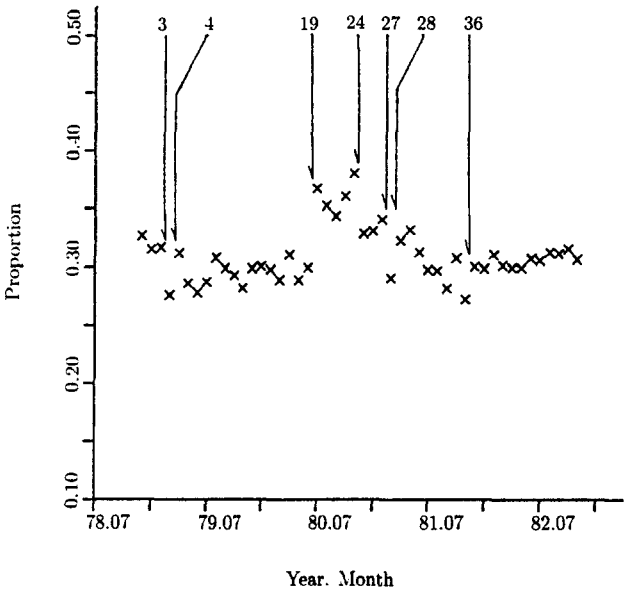

Fig. 2. A time-series plot of a public support rate for the Japan Liberal Democratic Party.

Table 4. Posterior probabilities of up to nine changes.

n	$p(N = n \mid \boldsymbol{y})$
0	0.000
1	0.001
2	0.059
3	0.237
4	0.192
5	0.154
6	0.113
7	0.098
8	0.082
9	0.064
Mean	5.0
Mode	3
Median	5

Table 4 presents the estimates of the posterior probabilities of up to nine changes as well as the posterior mean, mode and median of N. These results suggest that plural structural changes underlie the given series.

Table 5. Posterior probabilities of J_t's.

t	$p(J_t \mid \boldsymbol{y})$	t	$p(J_t \mid \boldsymbol{y})$
1	0.074	25	0.034
2	0.089	26	0.023
3	0.301	27	0.275
4	0.209	28	0.173
5	0.104	29	0.035
6	0.067	30	0.134
7	0.150	31	0.150
8	0.118	32	0.114
9	0.022	33	0.127
10	0.020	34	0.149
11	0.016	35	0.088
12	0.020	36	0.306
13	0.015	37	0.058
14	0.013	38	0.028
15	0.014	39	0.015
16	0.015	40	0.013
17	0.030	41	0.012
18	0.025	42	0.012
19	0.998	43	0.011
20	0.079	44	0.012
21	0.138	45	0.011
22	0.102	46	0.024
23	0.044	47	0.027
24	0.460		

Table 5 presents the estimate of each $p(J_t \mid \boldsymbol{y})$. The largest posterior probability is obtained at the 19th observation. Its value is almost equal to 1. This suggests that the 19th observation is a change point. Actually, it is widely recognized that the change in the opinion poll between June and July in 1980 was a remarkable one in the last two decades. This change is believed to have been caused by the sudden death of Prime Minister Oohira at the beginning of the election campaign that started in June 1980.

The second largest posterior probability is obtained at the 24th observation. Its value is not so large as the one at the 19th observation, but the 24th observation is also likely to be a change point since there are at least three change points according to the values of the mean, mode and median shown in Table 4. From Fig. 2, it is seen that the observed value largely shifts at the 24th observation.

Five other candidates for change points following the above two are at observation points 36, 3, 27, 4 and 28. Figure 2 shows that the changes at these points are prominent. The slope of the trend obviously changes at the 36th observation and the observed values largely shift at the other four points.

The Bayesian procedure identifies plural change points in the opinion poll data. These change points seem to agree with those views on the shifts of support for the LDP which were expressed by political observers and shown by analysis. The plot of the observation points also shows that these change points indicate the beginning of a shift in trend in the data.

Additionally, we note that the results obtained when repeating Step 1 seven times are very similar to the results given in Tables 4 and 5.

Acknowledgements

The author is grateful to the referees and the associate editor for their valuable comments.

REFERENCES

Akaike, H. (1977). On entropy maximization principle, *Applications of Statistics* (ed. P. R. Krishnaiah), 27–41, North-Holland, Amsterdam.

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

Bhattacharya, G. K. and Johnson, R. A. (1968). Nonparametric tests for shift at an unknown time point, *Ann. Math. Statist.*, **39**, 1731–1743.

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

Box, G. E. P. and Tiao, G. C. (1965). A change in level of a non-stationary time series, *Biometrika*, **52**, 181–192.

Broemeling, L. D. and Tsurumi, H. (1987). *Econometrics and Structural Change*, Dekker, New York.

Carlstein, E. (1988). Nonparametric change-point estimation, *Ann. Statist.*, **16**, 188–197.

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

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting, *J. Roy. Statist. Soc. Ser. B*, **38**, 205–247.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression, *Biometrika*, **56**, 495–504.

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

Hinkley, D. V. (1971). Inference in two-phase regression, *J. Amer. Statist. Assoc.*, **66**, 736–743.

James, B., James, K. L. and Siegmund, D. (1987). Tests for a change-point, *Biometrika*, **74**, 71–83.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series, *J. Amer. Statist. Assoc.*, **82**, 1032–1063.

Kitagawa, G. and Akaike, H. (1982). A quasi Bayesian to outlier detection, *Ann. Inst. Statist. Math.*, **34**, 389–398.

Lombard, F. (1987). Rank tests for changepoint problems, *Biometrika*, **74**, 615–624.

Page, E. S. (1954). Continuous inspection schemes, *Biometrika*, **41**, 100–114.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem, *Appl. Statist.*, **28**, 126–135.

Poirier, D. J. (1976). *The Econometrics of Structural Changes*, North-Holland, Amsterdam.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes, *J. Amer. Statist. Assoc.*, **53**, 873–880.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes, *J. Amer. Statist. Assoc.*, **55**, 324–330.

Schechtman, E. and Wolfe, D. A. (1985). Multiple change points problem—nonparametric procedures for estimation of the points of change, *Comm. Statist. B—Simulation Comput.*, **14**, 615–631.

Silvey, S. D. (1958). The Lindisfarne scribes' problem, *J. Roy. Statist. Soc. Ser. B*, **20**, 93–101.

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

Smith, A. F. M. (1980). Change-point problems: approaches and applications, *Trabajos Estadíst. Investigación Oper.*, **31**, 83–98.

Tanabe, K. and Tanaka, T. (1983). Fitting curves and surfaces by Bayesian models, *Chikyu*, **5**, 179–186 (in Japanese).

Tsurumi, H., Wago, H. and Ilmakunnas, P. (1986). Gradual switching multivariate regression models with stochastic cross-equational constraints and an application to the KLEM translog production model, *J. Econometrics*, **31**, 235–253.

Zacks, S. (1983). Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation, *Recent Advances in Statistics* (eds. M. H. Rizvi, J. S. Rustagi and D. O. Siegmund), 245–269, Academic Press, New York.