Ann. Inst. Statist. Math.
Vol. 45, No. 3, 511–530 (1993)

MAXIMUM LIKELIHOOD ESTIMATION IN THE MULTI-PATH CHANGE-POINT PROBLEM*

LAWRENCE JOSEPH[1]** AND DAVID B. WOLFSON[2]

[1] *Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, Canada H3G 1A4*
[2] *Department of Mathematics and Statistics, Burnside Hall, 805 Sherbrooke Street West, McGill University, Montreal, Quebec, Canada H3A 2K6*

(Received November 20, 1991; revised June 19, 1992)

Abstract. Maximum likelihood estimators of the parameters of the distributions before and after the change and the distribution of the time to change in the multi-path change-point problem are derived and shown to be consistent. The maximization of the likelihood can be carried out by using either the EM algorithm or results from mixture distributions. In fact, these two approaches give equivalent algorithms. Simulations to evaluate the performance of the maximum likelihood estimators under practical conditions, and two examples using data on highway fatalities in the United States, and on the health effects of urea formaldehyde foam insulation, are also provided.

Key words and phrases: Change-point, maximum likelihood estimation, EM algorithm, mixture distribution.

1. Introduction

The single-path change-point problem has been the subject of considerable research (Hinkley (1970), Cobb (1978), Shaban (1980), Picard (1985), Worsley (1986), Carlstein (1988)), while its counterpart, when the data consist of several sample paths, has received scant attention (Joseph (1989), Joseph and Wolfson (1992)). This is rather surprising, as many applications of "the change-point" arise when repeated observations are made in time, on different patients say, and it is desired to make inference about the instant of change (if any) in the health status of these patients.

Many of the techniques proposed for single-path change-point problems transfer, with suitable modification, to the multi-path setting. The method of maximum

* This work was supported in part by the Natural Science and Engineering Council of Canada, and the Fonds pour la Formation de chercheurs et l'aide à la Recherche Gouvernment du Québec.

** Lawrence Joseph is also a member of the Department of Epidemiology and Biostatistics of McGill University.

likelihood, for example, may be carried out by maximizing the joint likelihood of all the data with respect to both the unknown parameters of the underlying distributions of the data, as well as with respect to the unknown distribution of the position of change; it is assumed that each path has its own change-point and that there is a probability mass function describing the distribution of these change-points.

In the context of a single sample path it is well known that the maximum likelihood estimator of the change-point is not consistent as the number of observations on either side of the change-point tends to infinity, Hinkley (1970). This lack of consistency arises because the change-point problem is really a location problem on the space of infinite sequences and one finite path segment contains even less information than one observation in this space. Carlstein (1988) establishes consistency of a non-parametric change-point estimator of the ratio of the number of variables before to the number after the change in the single-path context.

It is the object here to discuss maximum likelihood estimation in a multi-path setting. An example on the change in traffic accident death rates after the relaxation of the 55 miles per hour speed limit in the U.S. in 1987, is provided as one illustration of the method of maximum likelihood. A second example concerning the health effects of urea formaldehyde foam insulation as measured by changes in the rate of visits to a doctor before and after installation of the material in the homes of a study population in Canada is also provided.

The general setup is given in Section 2, and the various approaches to maximum likelihood estimation are outlined in Section 3. It is soon realized that implementation of the maximum likelihood procedure in the multi-path change-point problem requires an approach more efficient than a point-by-point search. Section 4 discusses two of these approaches, one using the EM algorithm of Dempster *et al.* (1977), and the other taking a mixture viewpoint. In fact, these give equivalent algorithms. In anticipation of the examples in Section 7, consistency is proved for the Poisson case in Theorem 5.1. A sketch of a proof of consistency in the location-scale case is also given in Section 5. Section 6 presents the results of simulations designed to test the methods under a variety of practical circumstances, and Section 8 contains some concluding remarks.

The method of proof of Theorem 5.1 is based on the work of Kiefer and Wolfowitz (1956), who also point out that consistency of maximum likelihood estimators in a variety of multi-path settings often does not hold. However, when certain reasonable restrictions on the model are imposed and regularity conditions met, it is possible to establish consistency.

By regarding each path as arising from a mixture of distributions, consistency as well as asymptotic normality of all parameter estimates including those of the mixing constants, may be established in much the same way in the Poisson case, as has been done in the multivariate normal setting (Peters and Walker (1978)). The proof given there, however, requires assumptions that appear to be very difficult to check. For example, the information matrix arising from equation (3.1) below must be positive definite, which seems to be a daunting hurdle. There are virtually no general results for information matrices even in exponential family settings (see e.g. Redner and Walker (1984), p. 205). The compactness argument used to prove

consistency in this paper is straightforward and is related to the work of Redner (1981), Kiefer and Wolfowitz (1956) and Wald (1949).

Apart from the mathematical and statistical difficulties referred to in the above paragraph, there are also well known pitfalls in the actual solution of the likelihood equations (Redner (1984)). The EM algorithm offers a tractable solution. Here, simulations are used to verify the validity of the results. Application of general theorems as a means of verification appear to be very difficult if not impossible because of the complicated form of the likelihood function.

2. The general setup

We shall assume that the observations are in the form of the $M \times N$ array,

$$\begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix}. \tag{2.1}$$

A change will be said to occur at τ_i, in row i for $i = 1, 2, \dots, M$, and for $1 \leq \tau_i \leq N-1$, if $X_{i1}, X_{i2}, \dots, X_{i\tau_i}$, are identically distributed with common distribution F_1 which is different from the common distribution, F_2 of $X_{i\tau_i+1}, X_{i\tau_i+2}, \dots, X_{iN}$. In this paper our assumptions will be more specific in order to keep the exposition as simple as possible.

We make the following additional assumptions:

(i) The observations $\{X_{ij}\}$ within row i $(i = 1, 2, \dots, M)$, are independent $Poisson(\lambda_1)$ random variables for $j = 1, 2, \dots, \tau_i$ and independent $Poisson(\lambda_2)$ random variable's for $j = \tau_i + 1, \dots, N$ where $1 \leq \tau_i \leq N-1$. We say that no change has occurred in row i, if $\tau_i = N$. To simplify the proof of Theorem 5.1, we shall always assume that a change has occurred, although this assumption is not necessary.

(ii) The collection of rows in array (2.1) form independent random vectors.

(iii) The sequence $\{\tau_i\}_{i=1,2,\dots,M}$ is a set of independent and identically distributed discrete random variables with range the set of integers $\{1, 2, \dots, N-1\}$, and distribution function $G(\cdot)$ corresponding to the probability function $P_T(\tau) = \text{pr}(T = \tau)$.

(iv) The intensities, λ_1 and λ_2 are unknown, as is the distributional form of $G(\cdot)$.

In Section 3 maximum likelihood estimation of λ_1, λ_2 and $G(\cdot)$ will be discussed.

3. Maximum likelihood estimation

Given the set of observations $X_{ij} = x_{ij}$, $i = 1, 2, \dots, M$, $j = 1, 2, \dots, N$, the joint likelihood of the data in array (2.1) is given by

$$l(\vec{x}; \lambda_1, \lambda_2, \{P_T(\cdot)\}) = \prod_{i=1}^{M} \sum_{\tau_i=1}^{N-1} \prod_{j=1}^{\tau_i} f(x_{ij} \mid \lambda_1) \prod_{j=\tau_i+1}^{N} f(x_{ij} \mid \lambda_2) P_T(\tau_i), \tag{3.1}$$

where $f(\cdot \mid \lambda)$ denotes the Poisson probability function with mean λ.

The method of maximum likelihood calls for the maximization of $l(\vec{x}; \cdot)$ with respect to λ_1, λ_2, and the distribution $P_T(\tau)$. Alternative approaches to change-point inference including maximum likelihood methods under different assumptions, are given by Joseph (1989) and Joseph and Wolfson (1992).

There are at least three ways of interpreting $P_T(\tau)$ in this maximization problem:

(a) While the parameters λ_1 and λ_2 are fixed constants, the actual change points $\{\tau_i\}_{i=1,\ldots,M}$ are thought of as arising as realizations of M independent identically distributed random variables with unknown distribution $P_T(\cdot)$. This approach gives the proof of consistency presented in Section 5.

(b) The unknown true points of change $\{\tau_i\}$ are regarded as missing data and the ensuing maximization is carried out by using the EM algorithm.

(c) The quantities $P_T(\cdot)$ are looked upon as mixing constants in a standard finite mixture problem, where λ_1 and λ_2 are unknown parameters in the mixture distributions. It transpires that approaches (b) and (c) are equivalent, and is the justification for using the EM algorithm for mixture problems. See Section 4.

4. The EM algorithm and mixture distributions

If the τ_i's were observed, maximization of the likelihood with respect to λ_1 and λ_2 would, of course, be straightforward, as would be that of G, $\hat{G}$ being the empirical distribution function. Since the τ_i's are not observed we have a missing data problem where the missing data are the τ_i's . The EM algorithm, see Section 4.3 of Dempster *et al.* (1977), can then be applied as follows:

E-step: Given the current estimate at the k-th step, λ_1^k, λ_2^k and $P_T^k(\tau)$, estimate the $M \times N$ matrix, Z, of conditional probabilities that each row i has a change at position j, i.e.,

$$z_{ij} = \mathrm{pr}\{T_i = j \mid \lambda_1^k, \lambda_2^k, P_T^k(\tau), x\}$$

$$= \frac{\exp[-j\lambda_1^k - (N-j)\lambda_2^k + \log(\lambda_1^k)\sum_{l=1}^{j} x_{il} + \log(\lambda_2^k)\sum_{l=j+1}^{N} x_{il}] \times P_T^k(j)}{\sum_{j=1}^{N-1} \exp[-j\lambda_1^k - (N-j)\lambda_2^k + \log(\lambda_1^k)\sum_{l=1}^{j} x_{il} + \log(\lambda_2^k)\sum_{l=j+1}^{N} x_{il}] \times P_T^k(j)}$$

after some simplification, where x represents the array (2.1).

M-step: Compute

$$\lambda_1^{k+1} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(1-w_{ij})x_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N}(1-w_{ij})},$$

$$\lambda_2^{k+1} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}x_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}} \quad \text{and}$$

$$P_T^{k+1}(j) = \frac{\sum_{i=1}^{M} z_{ij}}{M},$$

where

$$w_{ij} = \begin{cases} 0, & j = 1, \text{ for all } i = 1, \ldots, M \\ w_{i,j-1} + z_{i,j-1}, & j = 2, \ldots, N, \text{ for all } i = 1, \ldots, M. \end{cases}$$

One alternates between the E-step and the M-step until a convergence criterion is met.

This same algorithm may be derived directly by maximizing the likelihood (3.1) using Lagrange multipliers with constraint $\sum_{j=1}^{N-1} P_T(j) = 1$. See, for example, Peters and Walker (1978), who address the problem of mixtures of multivariate normal distributions. On the other hand Redner and Walker (1984) derive their equations (4.5) and (4.6) from the EM algorithm. Here, the joint distribution of the random variables in each row is a mixture of exponential family random variables. This follows from the fact that the product of exponential family densities is again a member of an exponential family. Theorems 5.1 and 5.2 of Redner and Walker regarding convergence rates of the EM algorithm hold. In particular, under the Conditions 1 and 2 of Redner and Walker, p. 211, the algorithm converges to the unique asymptotically normal consistent maximum likelihood estimate if one can be shown to exist. Because the Conditions 1 and 2 of Redner and Walker are very difficult to check in the mixture context, we have performed simulations to examine the convergence of the maximum likelihood estimators computed via the EM algorithm. Our experience with the change-point problem is that if the initial parameter estimates are carefully chosen as described in Section 6, convergence to a global maximum always occurs. Consistency is the subject of Section 5.

5. Consistency

Our goal here is to show that given the array of data (2.1), the maximum likelihood estimators $\hat{G}$ and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are consistent as $M \to \infty$. Specifically, we have

THEOREM 5.1. *Under the data array* (2.1) *and under Assumptions* (i) *through* (iv) *above, the maximum likelihood estimators* $\hat{\lambda_1}$ *and* $\hat{\lambda_2}$ *satisfy*

$$\hat{\lambda_1} \to \lambda_1, \qquad \hat{\lambda_2} \to \lambda_2 \qquad \textit{almost surely as } M \to \infty,$$

while the maximum likelihood estimator $\hat{G}$ *of* G *satisfies*

$$\hat{G}(x) \to G(x) \qquad \textit{almost surely for each real } x \textit{ as } M \to \infty. \tag{5.1}$$

It was first pointed out by Neyman and Scott (1948) in a more general setting, that the maximum likelihood estimator of a parameter in the presence of infinitely many incidental parameters may not be consistent. That the situation changes when the incidental parameters are allowed to be independent and identically distributed random variables, is discussed by Kiefer and Wolfowitz (1956).

The setting of the Kiefer and Wolfowitz paper is that of an array of random variables $\{X_{ij}\}$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$ such that the density (mass) function of $X_{i1}, \ldots, X_{ik}$ is $f(\vec{x} \mid \vec{\theta}, \alpha_i)$ when $\vec{\theta}$ and α_i are given. Typically, $\vec{\theta}$ and α_i are

unknown and often it is desired to estimate $\vec{\theta}$, a so-called structural parameter, in the presence of the so-called incidental parameters $\{\alpha_i\}$. Kiefer and Wolfowitz show that under appropriate assumptions, the maximum likelihood estimator of $\vec{\theta}$ is consistent, provided $\{\alpha_i\}_i$ can be regarded as independent and identically distributed random variables with common distribution G. A bonus of their proof is that the nonparametric maximum likelihood estimator $\hat{G}$ of G is also consistent for G, an unexpected, and particularly useful result in the change-point setting.

Here, the data are the array (2.1), the parameter $\vec{\theta} = (\lambda_1, \lambda_2)$ and the incidental parameters $\alpha_i = \tau_i$, $i = 1, 2, \ldots, M$. Of course, in the usual change-point problem the distribution, G, of the τ_i, is of prime interest and the parameters λ_1 and λ_2 of secondary importance.

Our Assumptions 1–5 below are those of Kiefer and Wolfowitz, adapted to the change-point setup. Each is verified under the hypotheses of Section 2. It is important to notice that in the present change-point setting, $\hat{G}$ is termed nonparametric only because no parametric form is assumed; there are, of course, only finitely many points $G(\tau)$, $\tau = 1, 2, \ldots, N-1$ to estimate. Nevertheless, the work of Kiefer and Wolfowitz, most useful when there are infinitely many "incidental" parameters, facilitates a relatively straightforward proof of consistency in the present finite dimensional problem. In contrast, the well known Theorem 3.1 of Redner and Walker (1984), while establishing asymptotic normality as well as consistency, requires the positive definiteness of the information matrix, which is very difficult to compute in mixture problems. Related to the compactness approach of Kiefer and Wolfowitz is that of Redner (1981) whose results are intended for situations in which identifiability fails to hold. Equivalence classes of distributions replace distributions and consistency has to be interpreted with this in mind. Here, as Assumption 4 below shows, we do have identifiability, and this obviates the need to use Redner's result.

The following notation will be used:

The joint density of $X_{i1}, X_{i2}, \ldots, X_{iN}$ at the point $\vec{x}$, is denoted by $f(\vec{x} \mid \lambda_1, \lambda_2, \tau_i)$ when λ_1, λ_2, and τ_i are given. Of course,

$$f(\vec{x} \mid \lambda_1, \lambda_2, \tau_i) = \prod_{j=1}^{\tau_i} f(x_{ij} \mid \lambda_1) \prod_{j=\tau_i+1}^{N} f(x_{ij} \mid \lambda_2).$$

The "true" parameters and distribution respectively, will be denoted by ${}_0\lambda_1$, ${}_0\lambda_2$, and G_0. By definition, $\vec{\gamma} = (\lambda_1, \lambda_2, G)$, $\vec{\theta} = (\lambda_1, \lambda_2)$, and ${}_0\vec{\theta} = ({}_0\lambda_1, {}_0\lambda_2)$.

Let Ω be the space of possible values of $\vec{\theta}$, and $A = \{1, 2, \ldots, N-1\}$, the set of possible values of τ. The parameters $\theta_t^{(s)}$ $(1 \leq s \leq 2)$ will denote the components of a point $\vec{\theta}_t$ in Ω.

Let $\Gamma = \{G\}$ be a given space of cumulative distribution functions of τ_i. The G's will be discrete with support the finite set $\{1, 2, \ldots, N-1\}$. It is assumed that ${}_0\theta \in \Omega$, $G_0 \in \Gamma$, and $\vec{\gamma} \in \Omega \times \Gamma$.

We define

$$f(\vec{x} \mid \vec{\gamma}) = \int_A f(\vec{x} \mid \theta, \tau) dG(\tau) \tag{5.2}$$

$$= \sum_{\tau=1}^{N-1} \left\{ \prod_{j=1}^{\tau} f(x_j \mid \lambda_1) \prod_{j=\tau+1}^{N} f(x_j \mid \lambda_2) P_T(\tau) \right\},$$

where $P_T(N) = 0$.

In the space $\Omega \times \Gamma$, define the metric

$$\delta(\vec{\gamma_1}, \vec{\gamma_2}) = \sum_{s=1}^{2} |\arctan \vec{\theta}_1^{(s)} - \arctan \vec{\theta}_2^{(s)}| + \left\{ \sum_{i=1}^{N} |P_{T_1}(i) - P_{T_2}(i)|^2 \right\}^{1/2}, \tag{5.3}$$

and observe that $\delta(\vec{\gamma_1}, \vec{\gamma_2})$ converges to zero if and only if the first term on the r.h.s. of (5.3) and $|P_{T_1}(i) - P_{T_2}(i)|$ both converge to zero for $i = 1, 2, \ldots, N$. The metric (5.3), is equivalent to the metric on $\Omega \times \Gamma$ defined by Kiefer and Wolfowitz. The simplification arises because G is discrete, while the choice of metric permits compactification of $\Omega \times \Gamma$ to $\overline{\Omega} \times \overline{\Gamma}$. It should be noted that the additional functions needed to compactify $\Omega \times \Gamma$ need not be density functions.

PROOF OF THEOREM 5.1. The proof of Theorem 5.1 rests on a careful verification of Assumptions 1 through 5 below, in the multi-path change-point setting. These assumptions are discussed more generally by Kiefer and Wolfowitz (1956).

ASSUMPTION 1. $f(\vec{x} \mid \lambda_1, \lambda_2, \tau)$ is absolutely continuous with respect to a σ-finite measure on a Euclidean space of which $\vec{x}$ is a generic point.

Verification. This is obvious, f being a product of Poisson probability functions.

ASSUMPTION 2. Before checking Assumption 2, we shall need to "compactify" the space $\Omega \times \Gamma$. First, since $G \in \Gamma$ is discrete with range $\{1, 2, \ldots, N-1\}$, Γ is easily seen to be sequentially compact (in the sense of the metric defined by the second term on the right of (5.2)) and hence compact. Alternatively, if $\overline{\Gamma}$ is defined as the space Γ together with all the limits of its Cauchy sequences, then the completeness of $\mathcal{R}$ implies that $\overline{\Gamma} = \Gamma$.

Next, let $\overline{\Omega}$ be the space Ω together with the limits of all its Cauchy sequences (in the sense of the metric defined by the first term on the left hand side of (5.2)). Again, it is not difficult to show that if the points $(+\infty, +\infty)$, $(0, +\infty)$, $(+\infty, 0)$, and $(0, 0)$ are appended to Ω then we obtain $\overline{\Omega}$. The space $\overline{\Omega}$ is compact, and hence $\overline{\Omega} \times \Gamma$ is compact.

Returning to Assumption 2, let $\{\vec{\gamma}_n\}$ and $\{\vec{\gamma^*}\}$ belong to $\overline{\Omega} \times \Gamma$ and suppose that $\vec{\gamma}_n \to \vec{\gamma^*}$. We must show that $f(\vec{x} \mid \vec{\gamma}_n) \to f(\vec{x} \mid \vec{\gamma}^*)$, except perhaps on a set of $\vec{x}$ whose probability is 0 according to the density $f(\vec{x}|\vec{\gamma_0})$.

We complete the definition of f for $(\vec{\theta}, \tau)$ in $\overline{\Gamma} \times A$ in a straightforward manner: Define for all $\tau \in A$,

$$f(\vec{x} \mid \vec{\theta}, \tau) = \begin{cases} 0, & \text{if } x_i \neq 0, \text{ for some } i \leq \tau \text{ and if } \lambda_1 = 0 \\ 0, & \text{if } x_j \neq 0, \text{ for some } j \geq \tau \text{ and if } \lambda_2 = 0 \\ 0, & \text{if either } \lambda_1 = +\infty \text{ or if } \lambda_2 = +\infty, \text{ for all } \vec{x} \\ 1, & \text{if } \vec{x} = 0, \text{ and if } \lambda_1 = 0 \text{ and } \lambda_2 = 0 \\ e^{-\tau\lambda_1}, & \text{if } \vec{x} = 0, \text{ and if } \lambda_1 \neq 0, \lambda_2 = 0 \\ e^{-(N-\tau)\lambda_2}, & \text{if } \vec{x} = 0, \text{ and if } \lambda_1 = 0, \lambda_2 \neq 0. \end{cases}$$

For $(\vec{\theta}, G) \in \overline{\Omega} \times \overline{\Gamma}$, we then define $f(\vec{x} \mid \vec{\theta}, G)$ by (5.2).

Verification of Assumption 2. First suppose that $\vec{\gamma_n}$ and $\vec{\gamma}^*$ both $\in \overline{\Omega} \times \Gamma$, that $\vec{\gamma_n} \longrightarrow \vec{\gamma}^*$ and that at least one of λ_1^* and λ_2^* is infinite.

For instance, this may occur if the sequences $\{{}_n\lambda_1\}$ and $\{{}_n\lambda_2\}$ satisfy

$$ {}_n\lambda_1 \to \infty, \qquad {}_n\lambda_2 \to \lambda_2^* \in (0, +\infty) \qquad \text{as} \quad n \to \infty$$

i.e., $\vec{\gamma^*} = (+\infty, \lambda_2^*, G^*)$.

Then, by definition

$$f(\vec{x} \mid \vec{\gamma^*}) = \sum_{\tau=1}^{N-1} \prod_{i=1}^{\tau} f(x_i \mid \lambda_1^*) \prod_{i=\tau+1}^{N} f(x_i \mid \lambda_2^*) P_T(\tau) = 0.$$

On the other hand,

$$f(\vec{x} \mid \vec{\gamma_n}) = \sum_{\tau=1}^{N} \prod_{i=1}^{\tau} f(x_i \mid {}_n\lambda_1) \prod_{i=\tau+1}^{N} f(x_i \mid {}_n\lambda_2) P_{T_n}(\tau) \to 0$$

for all x. When ${}_n\lambda_1$ and ${}_n\lambda_2$ converge to λ_1^* and λ_2^* respectively, both nonzero and finite and $P_{T_n} \to P_T$ the continuity of $f(\cdot \mid \lambda_1)$ and $f(\cdot \mid \lambda_2)$ in the arguments λ_1, and λ_2 respectively, ensures, trivially, that $f(\vec{x} \mid \vec{\gamma_n}) \to f(\vec{x} \mid \vec{\gamma}^*)$.

ASSUMPTION 3. For any $\vec{\gamma} \in \overline{\Omega} \times \Gamma$, and any $\rho > 0$, $w(\vec{x} \mid \vec{\gamma}, \rho)$ is a measurable function of $\vec{x}$ where $w(\vec{x} \mid \vec{\gamma}, \rho) = \sup f(\vec{x} \mid \vec{\gamma}')$, the sup being taken over all $\vec{\gamma}' \in \overline{\Omega} \times \Gamma$ for which $\delta(\vec{\gamma}, \vec{\gamma}') < \rho$.

Verification. The measurability of $w(\vec{x} \mid \vec{\gamma}, \rho)$ follows immediately from the measurability of $f(\vec{x} \mid \vec{\gamma}')$.

ASSUMPTION 4. (Identifiability) If $\vec{\gamma}_1 \in \overline{\Omega} \times \Gamma$ is different from $\vec{\gamma}_0$, then, for at least one y,

$$\int_{-\infty}^{y} \cdots \int_{-\infty}^{y} f(\vec{x} \mid \vec{\gamma}_1) d\vec{x} \neq \int_{-\infty}^{y} \cdots \int_{-\infty}^{y} f(\vec{x} \mid \vec{\gamma}_0) d\vec{x}.$$

Verification. Since by assumption, the "true parameter" $\vec{\gamma_0} \in \Omega \times \Gamma$, it is sufficient to show that if, for almost all $\vec{x}$ and $\vec{\gamma_1} \in \overline{\Omega} \times \overline{\Gamma}$

$$f(\vec{x} \mid \vec{\gamma}_1) = f(\vec{x} \mid \vec{\gamma}_0), \tag{5.4}$$

then $\vec{\gamma}_1 = \vec{\gamma}_0$. Now, (5.4) implies

$$\begin{aligned} \sum_{\tau=1}^{N-1} \left\{ \prod_{i=1}^{\tau} f(x_i \mid {}_1\lambda_1) \prod_{i=\tau+1}^{N} f(x_i \mid {}_1\lambda_2) P_{T_1}(\tau) \right\} \\ = \sum_{\tau=1}^{N-1} \left\{ \prod_{i=1}^{\tau} f(x_i \mid {}_0\lambda_1) \prod_{i=\tau+1}^{N} f(x_i \mid {}_0\lambda_2) P_{T_0}(\tau) \right\} \end{aligned} \tag{5.5}$$

for almost all $\vec{x}$.

Summing both sides over $x_2, x_3, \ldots, x_N$ we get

$$\sum_{\tau=1}^{N-1} f(x_1 \mid {}_1\lambda_1) P_{T_1}(\tau) = \sum_{\tau=1}^{N-1} f(x_1 \mid {}_0\lambda_1) P_{T_0}(\tau)$$

$$\Rightarrow f(x_1 \mid {}_1\lambda_1) \sum_{\tau=1}^{N-1} P_{T_1}(\tau) = f(x_1 \mid {}_0\lambda_1) \sum_{\tau=1}^{N-1} P_{T_0}(\tau)$$

$$\Rightarrow f(x_1 \mid {}_1\lambda_1) = f(x_1 \mid {}_0\lambda_1)$$

for all x_1. But this implies that ${}_1\lambda_1 = {}_0\lambda_1$.

Similarly, by summing over $x_1, x_2, \ldots, x_{N-1}$, we get ${}_1\lambda_2 = {}_0\lambda_2$, and hence we obtain $({}_1\lambda_1, {}_1\lambda_2) = ({}_0\lambda_1, {}_0\lambda_2)$.

To prove that $P_{T_1}(\tau) = P_{T_0}(\tau)$ for $\tau = 1, 2, \ldots, N-1$, sum both sides of (5.4) over $x_3, x_4, \ldots, x_{N-1}$. We get

$$\begin{aligned}
&f(x_1 \mid {}_1\lambda_1) f(x_2 \mid {}_1\lambda_2) P_{T_1}(1) + f(x_1 \mid {}_1\lambda_1) f(x_2 \mid {}_1\lambda_1) P_{T_1}(2) \\
&\quad + \cdots + f(x_1 \mid {}_1\lambda_1) f(x_2 \mid {}_1\lambda_1) P_{T_1}(N-1) \\
&= f(x_1 \mid {}_0\lambda_1) f(x_2 \mid {}_0\lambda_2) P_{T_0}(1) + f(x_1 \mid {}_0\lambda_1) f(x_2 \mid {}_0\lambda_1) P_{T_0}(2) \\
&\quad + \cdots + f(x_1 \mid {}_0\lambda_1) f(x_2 \mid {}_0\lambda_1,) P_{T_0}(N-1), \qquad \text{i.e.,} \\
&f(x_1 \mid {}_1\lambda_1)[f(x_2 \mid {}_1\lambda_2) P_{T_1}(1) + f(x_2 \mid {}_1\lambda_1)\{1 - P_{T_1}(1)\}] \\
&\quad = f(x_1 \mid {}_0\lambda_1)[f(x_2 \mid {}_0\lambda_2,) P_{T_0}(1) + f(x_2 \mid {}_0\lambda_1)\{1 - P_{T_0}(1)\}].
\end{aligned}$$

But we have already established $({}_1\lambda_1, {}_1\lambda_2) = ({}_0\lambda_1, {}_0\lambda_2) = (\lambda_1, \lambda_2)$, say. Therefore, (5.4) implies that

$$f(x_2 \mid \lambda_2) P_{T_1}(1) - f(x_2 \mid \lambda_1) P_{T_1}(1) = f(x_2 \mid \lambda_2) P_{T_0}(1) - f(x_2 \mid \lambda_1) P_{T_0}(1),$$

i.e.,

$$\begin{aligned}
P_{T_1}(1)[f(x_2 \mid \lambda_2) - f(x_2 \mid \lambda_1)] &= P_{T_0}(1)[f(x_2 \mid \lambda_2) - f(x_2 \mid \lambda_1)] \\
&\Rightarrow P_{T_1}(1) = P_{T_0}(1),
\end{aligned}$$

since $\lambda_1 \neq \lambda_2 \Rightarrow \exists x_2$ such that $f(x_2 \mid \lambda_1) \neq f(x_2 \mid \lambda_2)$.

In the same way, summing over $x_k, x_{k+1}, \ldots, x_{N-1}$, one can show iteratively that $P_{T_1}(k-2) = P_{T_0}(k-2)$, for $k = 3, \ldots, N-1$.

To obtain $P_{T_1}(N-1) = P_{T_0}(N-1)$ and hence $P_{T_1}(N-2) = P_{T_0}(N-2)$, we simply carry out the above procedure except we sum over $x_1, x_2, \ldots, x_{N-2}$. Hence, $P_{T_1}(k) = P_{T_0}(k)$ for all $k = 1, 2, \ldots, N-1$, and it follows that $\vec{\gamma_0} = \vec{\gamma_1}$.

Since a full inductive proof is cumbersome, it is omitted.

ASSUMPTION 5. (Integrability Assumption) For any $\vec{\gamma} \in \overline{\Omega} \times \Gamma$, we have, as $\rho \downarrow 0$,

$$\lim E\left[\log \frac{w(\vec{X} \mid \vec{\gamma}, \rho)}{f(\vec{X} \mid \vec{\gamma}_0)}\right]^+ < \infty.$$

$$\lim_{\rho\downarrow 0} E\left[\log \frac{w(\vec{X} \mid \vec{\gamma}, \rho)}{f(\vec{X} \mid \vec{\gamma}_0)}\right]^+ = \lim_{\rho\downarrow 0} E[\log w(\vec{X} \mid \vec{\gamma}, \rho) - \log f(\vec{X} \mid \vec{\gamma}_0)]^+$$
$$= \lim_{\rho\downarrow 0} E[\log \sup f(\vec{X} \mid \vec{\gamma}, \rho) - \log f(\vec{X} \mid \vec{\gamma}_0)]^+$$
$$\leq \lim_{\rho\downarrow 0} E|\log \sup f(\vec{X} \mid \vec{\gamma}, \rho)| + E|\log f(\vec{X} \mid \vec{\gamma}_0)|,$$

and Assumption 5 follows since $\vec{\gamma}_0 \in \Omega \times \Gamma$ and $f(\cdot)$ is a Poisson probability function.

The verification of Assumptions 1 to 5 implies that the maximum likelihood estimators of λ_1 and λ_2 are strongly consistent and, perhaps more importantly, the maximum likelihood estimator of G is consistent in the sense that for each x, $\hat{G}_n(x)$ converges almost surely to $G(x)$ at all continuity points x of G.

Generalization of Theorem 5.1. Consistency can, in fact, be established for a much wider class of underlying distributions than the Poisson. We state such a result for location-scale families and sketch the main arguments of the proof.

THEOREM 5.2. *Suppose that the joint likelihood of the data in array* (2.1) *is given by*

$$l(\vec{x}; \alpha_1, \beta_1, \alpha_2, \beta_2, \{P_T(\cdot)\})$$
$$= \prod_{i=1}^{M} \sum_{\tau_i=1}^{N-1} \prod_{j=1}^{\tau_i} \frac{1}{\beta_1} f\left(\frac{x_{ij} - \alpha_1}{\beta_1}\right) \prod_{j=\tau_i+1}^{N} \frac{1}{\beta_2} f\left(\frac{x_{ij} - \alpha_2}{\beta_2}\right) P_T(\tau_i),$$

where we assume that f is a density with respect to a σ-finite measure on a Euclidean k-space, and

(5.6) $\beta_1 \geq c > 0$ *and* $\beta_2 \geq c > 0$, c *known*,

(5.7) $\sup_x f(x) < \infty$,

(5.8) f *is a measurable function of* x,

(5.9) $\lim_{|x|\to\infty} f(x) = 0$, *and*

(5.10) $\int_{-\infty}^{\infty} f(x)[\log |x|]^+ dx < \infty$.

Then under assumptions (i) *through* (iv) *above (with the obvious modifications), the maximum likelihood estimators $\hat{\alpha_1}$, $\hat{\alpha_2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ satisfy*

$$\hat{\alpha_1} \to \alpha_1, \quad \hat{\alpha_2} \to \alpha_2, \quad \hat{\beta}_1 \to \beta_1, \quad \text{and} \quad \hat{\beta}_2 \to \beta_2$$

almost surely as $M \to \infty$, while the maximum likelihood estimator $\hat{G}$ of G satisfies $\hat{G} \to G$ almost surely for each real x as $M \to \infty$.

The conditions (5.6) to (5.10) are a slightly modified subset of those given by Kiefer and Wolfowitz ((1956), p. 895). The current situation is simpler than

theirs as G is a discrete distribution. The condition (5.6) is, in practice hardly restrictive, as one could use $c = 10^{-6}$, say. Many important distributions are covered by Theorem 5.2, including the normal, uniform, Cauchy, and exponential.

Sketch of the proof of Theorem 5.2. As in the proof of Theorem 5.1, Assumptions 1 to 5 must be verified. The salient features only, are given. Assumption 1 is trivial, and Assumption 2 follows by defining for each τ_i,

$$\prod_{j=1}^{\tau_i} \frac{1}{\beta_1} f\left(\frac{x_{ij} - \alpha_1}{\beta_1}\right) \prod_{j=\tau_i+1}^{N} \frac{1}{\beta_2} f\left(\frac{x_{ij} - \alpha_2}{\beta_2}\right) = 0$$

if $\alpha_1 = \pm\infty$ or $\alpha_2 = \pm\infty$, or $\beta_1 = +\infty$ or $\beta_2 = +\infty$, and then invoking conditions (5.7) and (5.9).

Assumption 3 is a consequence of condition (5.8).

Assumption 4 follows from the identifiability of location scale families: if $F(x) = F(ax + b)$ for all x and if F is non-degenerate, then $a = 1$ and $b = 0$. (For a proof of this result see Billingsley (1986).)

Assumption 5 is proved using the assumed boundedness from below of β_1 and β_2 and the uniform boundedness of $f(x)$, condition (5.7). The verification of Assumption 5 then involves a straightforward adaption of the verification of Assumption 5 in example 1a, p. 895, of Kiefer and Wolfowitz.

Apart from the Poisson case given by Theorem 5.1, the gamma distribution with density

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, & \text{for } x > 0 \\ 0, & \text{elsewhere} \end{cases}$$

is an important special case not covered by the location scale families of Theorem 5.2. Again, however, by imposing the harmless restriction $\alpha_1 \geq c > 0$, $\alpha_2 \geq c > 0$ on the shape parameters before and after the change, consistency of the maximum likelihood estimators continues to hold. Here, the change is in the vector (α, β).

It is possible to define $A = \{1, 2, \ldots, N\}$, in which case there is positive probability, $P_T(N)$, of no change. This situation involves no special difficulties in the proof of consistency, although for brevity of exposition the simpler case of $A = \{1, 2, \ldots, N-1\}$ is given. If $P_T(N) = 1$ then with probability one there is no change and the proof of identifiability breaks down. This shortcoming may, however, be avoided by simply assuming that $P_T(N) < 1$, which for practical purposes is not unreasonable.

While consistency provides a theoretical justification for the maximum likelihood estimators, simulations to assess the effectiveness of the EM algorithm and also to examine the feasibility of using the method for small samples are important.

6. Simulations

Referring to (2.1), fix M sequences of length $N = 8$ or $N = 40$, where $M = 10$, 30, 100, or 500. Two sets of choices for the Poisson parameters for F_1 and F_2 were considered: a change from $\lambda_1 = 3$ to $\lambda_2 = 5$, and a change from $\lambda_1 = 2$ to $\lambda_2 = 6$. These values were selected by choosing the lowest integers that solved the equation $(\lambda_1 - \lambda_2)/\sqrt{(\lambda_1 + \lambda_2)/2} = c$, for $c = 1$ or 2, corresponding to "standardized differences" for the change in mean of size 1 and 2, respectively.

For $N = 40$, five choices for P_T were considered:

1. $U(15, 24)$, a uniform distribution on the integers from 15 to 24, so that $\text{pr}\{\tau = k\} = 0.1$, $k = 15, \ldots, 24$, and zero elsewhere.

2. $U(1, 40)$, a uniform distribution on the integers from 1 to 40, so that $\text{pr}\{\tau = k\} = 0.025$, $k = 1, \ldots, 40$.

3. $T(15, 25)$, a "tent-shaped" function, with peak at $\tau = 20$, and sloping down linearly to zero at $\tau = 15$ and $\tau = 25$. Hence $\text{pr}\{\tau = 20\} = 0.2$, $\text{pr}\{\tau = 19\} = \text{pr}\{\tau = 21\} = 0.16$, $\text{pr}\{\tau = 18\} = \text{pr}\{\tau = 22\} = 0.12$, $\text{pr}\{\tau = 17\} = \text{pr}\{\tau = 23\} = 0.08$, $\text{pr}\{\tau = 16\} = \text{pr}\{\tau = 24\} = 0.04$, all other choices for k having zero probability.

4. $T(3, 7)$, a tent-shaped function, with peak at $\tau = 5$, and sloping down linearly to zero at $\tau = 3$ and $\tau = 7$. Hence $\text{pr}\{\tau = 5\} = 0.5$, and $\text{pr}\{\tau = 4\} = \text{pr}\{\tau = 6\} = 0.25$, all other choices for k having zero probability.

5. $S(4, 38)$, a spiked function, where $\text{pr}\{\tau = 4\} = \text{pr}\{\tau = 38\} = 0.5$ all other choices for k having zero probability.

Using similar notation, the choices of P_T for $N = 8$ were $U(1, 8)$, $T(3, 7)$, and $S(2, 6)$.

These choices cover a wide range of possible shapes for the distributions of τ. Included are those where nothing at all is known about the location of the change, $U(1, 40)$ and $U(1, 8)$, distributions where the change is equally likely to occur in a specified region, but nothing is known about the relative probabilities within the region, $U(15, 25)$, two distributions where the most likely location for the change is known, but points near this value are also possible, with decreasing probability further from the centre, $T(15, 25)$ and $T(3, 7)$, and a distribution where the change may occur early or late in the sequence, with equal probability, $S(4, 38)$ and $S(2, 6)$.

These choices for P_T, combined with two sets of λ parameters and four choices for M gives 40 different situations for $N = 40$, and 24 for $N = 8$. Each of these 64 combinations was simulated 300 times.

Outcome measures for the simulations include the average error in P_T, defined by $|P_T(t) - \hat{P}_T(t)|$ averaged over all N possible locations for t and over all 300 simulations, and various statistics summarizing the largest error, defined by $\sup_{1 \le t \le N} |P_T(t) - \hat{P}_T(t)|$. These included the mean, median, and range of the largest error, over the 300 simulations. The stopping criterion was $\sup_{1 \le t \le N} |P_T^{k+1}(t) - P_T^k(t)| \le 0.00001$. For all outcome measures, $P_T(\cdot)$ was taken to be the empirical distribution function of the realized τ's. This was chosen rather than the theoretical $G(\tau)$ to make the simulations with small M meaningful.

The initial estimates of P_T were taken to be the empirical distribution function

Table 1. Results of the simulations, number of columns = 40.

#	M	$G(\tau)$	λ_1	λ_2	Mean	Median	Range	Avg error
1	10	$U(15,24)$	3	5	0.3298	0.3015	(0.1031, 0.8055)	0.0306
2	10	$U(1,40)$	3	5	0.2492	0.2353	(0.1000, 0.4808)	0.0339
3	10	$T(15,25)$	3	5	0.3392	0.3057	(0.1000, 0.9109)	0.0283
4	10	$T(3,7)$	3	5	0.3670	0.3561	(0.0329, 0.8371)	0.0232
5	10	$S(4,38)$	3	5	0.3193	0.3000	(0.0000, 0.7998)	0.0213
6	10	$U(15,24)$	2	6	0.1730	0.1761	(0.0207, 0.4922)	0.0157
7	10	$U(1,40)$	2	6	0.1351	0.1118	(0.0018, 0.2998)	0.0168
8	10	$T(15,25)$	2	6	0.1938	0.1985	(0.0600, 0.6863)	0.0154
9	10	$T(3,7)$	2	6	0.1901	0.1671	(0.0034, 0.5525)	0.0105
10	10	$S(4,38)$	2	6	0.0938	0.0873	(0.0000, 0.5000)	0.0055
11	30	$U(15,24)$	3	5	0.2149	0.2017	(0.0935, 0.4759)	0.0239
12	30	$U(1,40)$	3	5	0.1531	0.1494	(0.0708, 0.3984)	0.0309
13	30	$T(15,25)$	3	5	0.2352	0.2192	(0.0667, 0.6006)	0.0221
14	30	$T(3,7)$	3	5	0.2456	0.2320	(0.0377, 0.6324)	0.0153
15	30	$S(4,38)$	3	5	0.1733	0.1465	(0.0003, 0.5402)	0.0115
16	30	$U(15,24)$	2	6	0.1072	0.1000	(0.0256, 0.2732)	0.0112
17	30	$U(1,40)$	2	6	0.0773	0.0712	(0.0333, 0.1738)	0.0163
18	30	$T(15,25)$	2	6	0.1153	0.1096	(0.0359, 0.2678)	0.0106
19	30	$T(3,7)$	2	6	0.1004	0.0954	(0.0089, 0.2664)	0.0055
20	30	$S(4,38)$	2	6	0.0506	0.0404	(0.0000, 0.2471)	0.0031
21	100	$U(15,24)$	3	5	0.1421	0.1336	(0.0648, 0.4200)	0.0170
22	100	$U(1,40)$	3	5	0.0955	0.0910	(0.0500, 0.1851)	0.0249
23	100	$T(15,24)$	3	5	0.1503	0.1399	(0.0486, 0.4436)	0.0153
24	100	$T(3,7)$	3	5	0.1509	0.1428	(0.0183, 0.3706)	0.0089
25	100	$S(4,38)$	3	5	0.0883	0.0781	(0.0005, 0.2343)	0.0061
26	100	$U(15,24)$	2	6	0.0574	0.0562	(0.0226, 0.2731)	0.0062
27	100	$U(1,40)$	2	6	0.0420	0.0401	(0.0221, 0.0789)	0.0122
28	100	$T(15,24)$	2	6	0.0609	0.0571	(0.0172, 0.1372)	0.0059
29	100	$T(3,7)$	2	6	0.0586	0.0557	(0.0019, 0.2198)	0.0032
30	100	$S(4,38)$	2	6	0.0252	0.0208	(0.0001, 0.1158)	0.0016
31	500	$U(15,24)$	3	5	0.0751	0.0730	(0.0300, 0.1720)	0.0087
32	500	$U(1,40)$	3	5	0.0497	0.0479	(0.0297, 0.0960)	0.0161
33	500	$T(15,24)$	3	5	0.0755	0.0715	(0.0263, 0.1603)	0.0079
34	500	$T(3,7)$	3	5	0.0667	0.0625	(0.0071, 0.1699)	0.0041
35	500	$S(4,38)$	3	5	0.0397	0.0324	(0.0051, 0.1547)	0.0028
36	500	$U(15,24)$	2	6	0.0230	0.0226	(0.0099, 0.0417)	0.0026
37	500	$U(1,40)$	2	6	0.0185	0.0183	(0.0118, 0.0280)	0.0057
38	500	$T(15,24)$	2	6	0.0268	0.0256	(0.0098, 0.0602)	0.0026
39	500	$T(3,7)$	2	6	0.0241	0.0217	(0.0021, 0.0757)	0.0013
40	500	$S(4,38)$	2	6	0.0090	0.0077	(0.0001, 0.0343)	0.0006

Table 2. Results of the simulations, number of columns = 8.

#	M	$G(\tau)$	λ_1	λ_2	Mean	Median	Range	Avg error
41	10	$U(1,8)$	3	5	0.3785	0.3297	(0.1000, 0.9991)	0.1364
42	10	$T(3,7)$	3	5	0.3700	0.3278	(0.0441, 0.9073)	0.1090
43	10	$S(2,6)$	3	5	0.3739	0.3570	(0.0066, 0.9997)	0.1103
44	10	$U(1,8)$	2	6	0.1793	0.1812	(0.0107, 0.4000)	0.0709
45	10	$T(3,7)$	2	6	0.1922	0.1890	(0.0108, 0.7124)	0.0528
46	10	$S(2,6)$	2	6	0.1189	0.0981	(0.0000, 0.5145)	0.0336
47	30	$U(1,8)$	3	5	0.2429	0.2266	(0.0843, 0.5767)	0.1047
48	30	$T(3,7)$	3	5	0.2557	0.2475	(0.0365, 0.5461)	0.0752
49	30	$S(2,6)$	3	5	0.2265	0.2107	(0.0032, 0.6333)	0.0679
50	30	$U(1,8)$	2	6	0.1125	0.1028	(0.0201, 0.2882)	0.0486
51	30	$T(3,7)$	2	6	0.1076	0.1005	(0.0123, 0.2991)	0.0300
52	30	$S(2,6)$	2	6	0.0600	0.0481	(0.0004, 0.2328)	0.0172
53	100	$U(1,8)$	3	5	0.1515	0.1468	(0.0466, 0.3236)	0.0693
54	100	$T(3,7)$	3	5	0.1489	0.1435	(0.0169, 0.3727)	0.0446
55	100	$S(2,6)$	3	5	0.1268	0.1221	(0.0080, 0.3952)	0.0390
56	100	$U(1,8)$	2	6	0.0603	0.0560	(0.0193, 0.1353)	0.0276
57	100	$T(3,7)$	2	6	0.0559	0.0512	(0.0047, 0.1543)	0.0153
58	100	$S(2,6)$	2	6	0.0293	0.0259	(0.0002, 0.0994)	0.0085
59	500	$U(1,8)$	3	5	0.0782	0.0763	(0.0248, 0.1947)	0.0355
60	500	$T(3,7)$	3	5	0.0642	0.0596	(0.0119, 0.1972)	0.0195
61	500	$S(2,6)$	3	5	0.0585	0.0570	(0.0056, 0.1366)	0.0181
62	500	$U(1,8)$	2	6	0.0265	0.0258	(0.0095, 0.0564)	0.0123
63	500	$T(3,7)$	2	6	0.0255	0.0253	(0.0027, 0.0761)	0.0070
64	500	$S(2,6)$	2	6	0.0125	0.0111	(0.0001, 0.0355)	0.0037

of estimates from the single-path maximum likelihood estimator for a change, Hinkley (1970). The initial estimates of λ_1 and λ_2 were the weighted averages of observations before and after the change in each row.

In all cases, maximum likelihood estimates were computed by the EM algorithm as described in Section 4, programmed in Fortran, and run on a SPARCstation SLC.

Results of the simulations. The results of the simulations are tabulated in Tables 1 and 2, and boxplots comparing the distributions of the greatest errors appear in Figs. 1 and 2. The estimates for the λ's are not included in the tables. For $N = 40$, they were almost always accurately estimated. In fact, $|\hat{\lambda}_i - \lambda_i| \leq 0.26$ across all 40×300 simulations, $|\hat{\lambda}_i - \lambda_i| \leq 0.15$ for simulations with $M \geq 100$, and $|\hat{\lambda}_i - \lambda_i| \leq 0.05$ for simulations with $M = 500$, $i = 1, 2$. As could be expected, the accuracy for $N = 8$ was somewhat less than for $N = 40$, with $|\hat{\lambda}_i - \lambda_i| \leq 5$ across all 24×300 simulations, $|\hat{\lambda}_i - \lambda_i| \leq 0.67$ for simulations with $M \geq 100$, and $|\hat{\lambda}_i - \lambda_i| \leq 0.22$ for simulations with $M = 500$, $i = 1, 2$. However, in most simulations λ_1 and

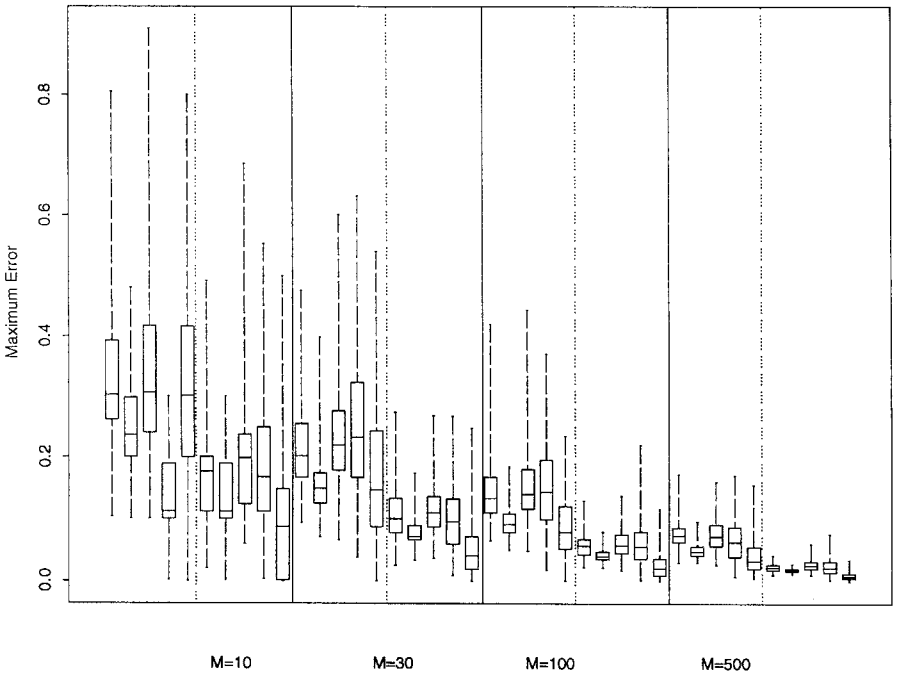

Fig. 1. Results of the simulations for $N = 40$: boxplots of the largest errors.

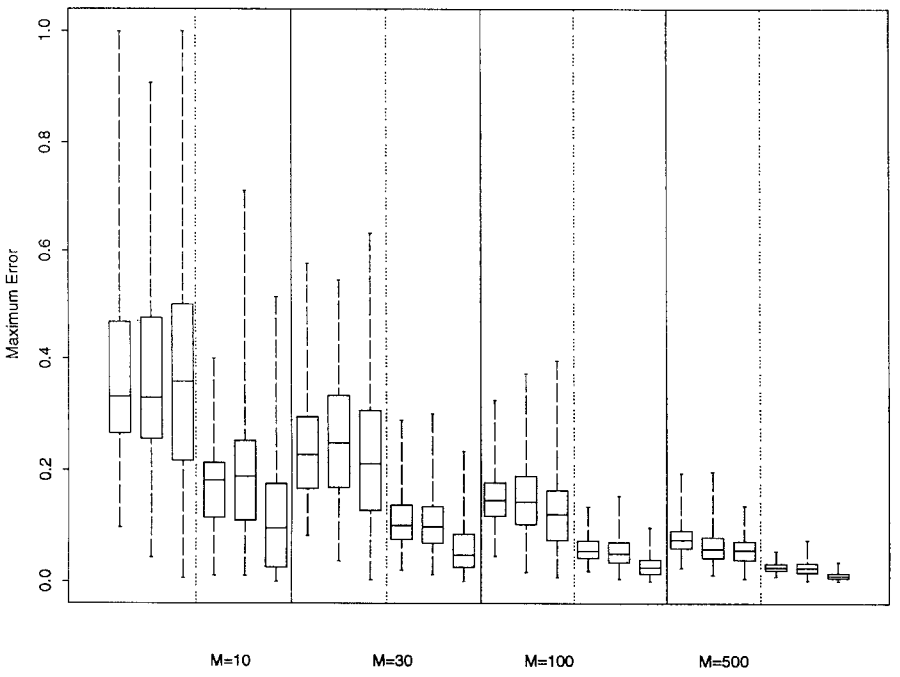

Fig. 2. Results of the simulations for $N = 8$: boxplots of the largest errors.

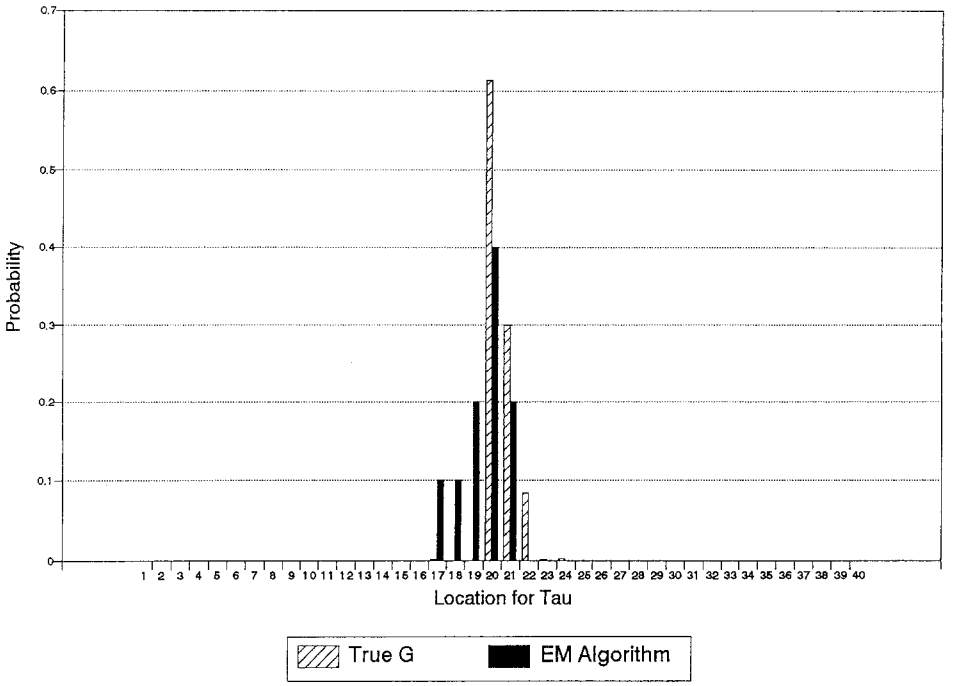

Fig. 3. Sample result of a simulation $M = 10$, $G = T(15, 25)$, $\lambda_1 = 3$, $\lambda_2 = 5$.

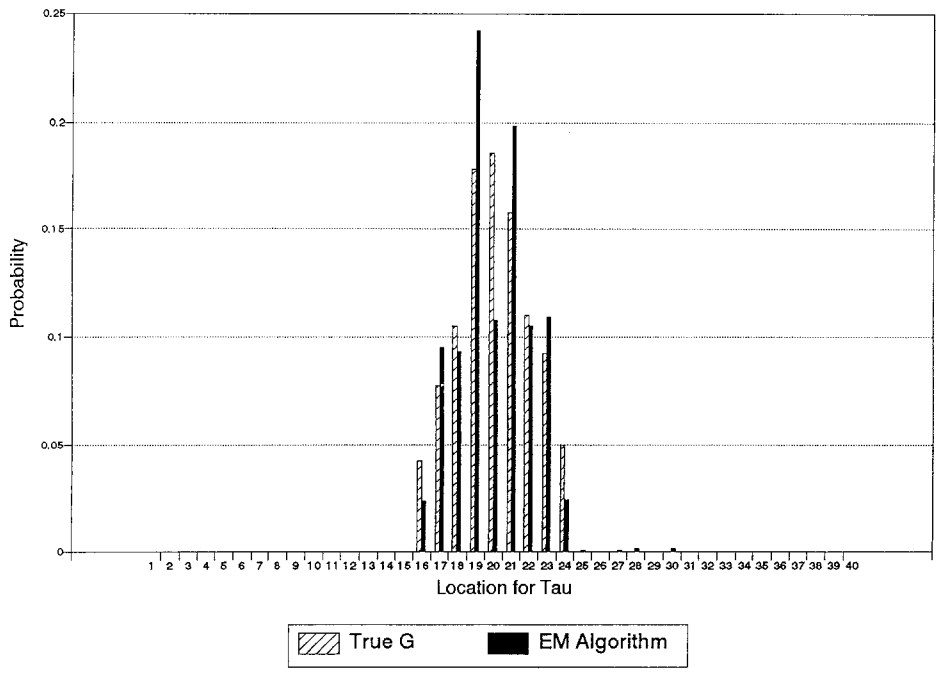

Fig. 4. Sample result of a simulation $M = 500$, $G = T(15, 25)$, $\lambda_1 = 3$, $\lambda_2 = 5$.

λ_2 were estimated with much more accuracy than these maximal errors indicate.

As expected, errors in estimation of the change-point distribution decreased appreciably as M increased. For $M = 500$ even the maximum error was typically much less than 0.08, while maximum errors of 0.3 were typical when $M = 10$. As Fig. 3 shows, however, even with $M = 10$, the estimated probability tended to be in the same neighbourhood as the true probability, but was often moved over by one or two indices along the x-axis. Under most circumstances, these errors should not greatly decrease the value of the analysis. Figure 4 shows the improvement for $M = 500$. There was great similarity between the cases where $N = 40$ and $N = 8$ in these measurements.

7. Examples

The analyses in this section while by no means complete serve to illustrate the method proposed in this paper.

Example 1. Urea-formaldehyde foam insulation (UFFI) was installed in many homes in Canada until it was banned by the Federal Government on December 18, 1980. The decision to ban UFFI was based more on precautionary measures than solid evidence that the material was harmful to the health of residents of the buildings in which it was installed. One indicator of the danger posed by UFFI would be an increase in the rate at which household occupants visit a doctor after installation as compared to before. Of course, even if there is an increase, one would not expect an instant reaction to the material, and one may be interested in estimating the time to effect (if any). Tri-monthly data was collected, L'Abbé (1984), on the number of visits to a doctor for one year before and after installation of the foam in 337 households in Canada, so that $N = 8$. There were 67 cases with missing data (an interesting extension of the methods presented here would be to employ the EM algorithm to simultaneously estimate the missing values in addition to the usual parameters of the model), and a further 36 outliers, defined as those cases with greater than 20 visits in any three month period, presumed to have a serious illness not associated with UFFI. After removing these cases, $M = 234$ households remained in the analysis.

The algorithm as proposed in Section 4 was employed. The visit rates were estimated to be $\hat{\lambda_1} = 1.12$ and $\hat{\lambda_2} = 4.83$. The results for $\hat{P}_T$ are given by Fig. 5, where it is noted that $\hat{P}(8) = 0.55$. This may be interpreted to mean that there is relatively high probability that UFFI had no effect on the rate at which most household occupants visited their doctors. Other research on this contentious issue supports these findings. This analysis required only 48 iterations and approximately 20 seconds to converge.

Example 2. Data consisting of the number of rural highway fatalities in the United States before and after the relaxation of the 55 mile per hour speed limit were analyzed. Precisely, the data provided by the United States National Highway Traffic Safety Administration gave the monthly number of traffic deaths on rural interstate roads for all states from April 1985 to April 1989. This time period included the year that the 55 miles per hour speed limit was lifted by Congress.

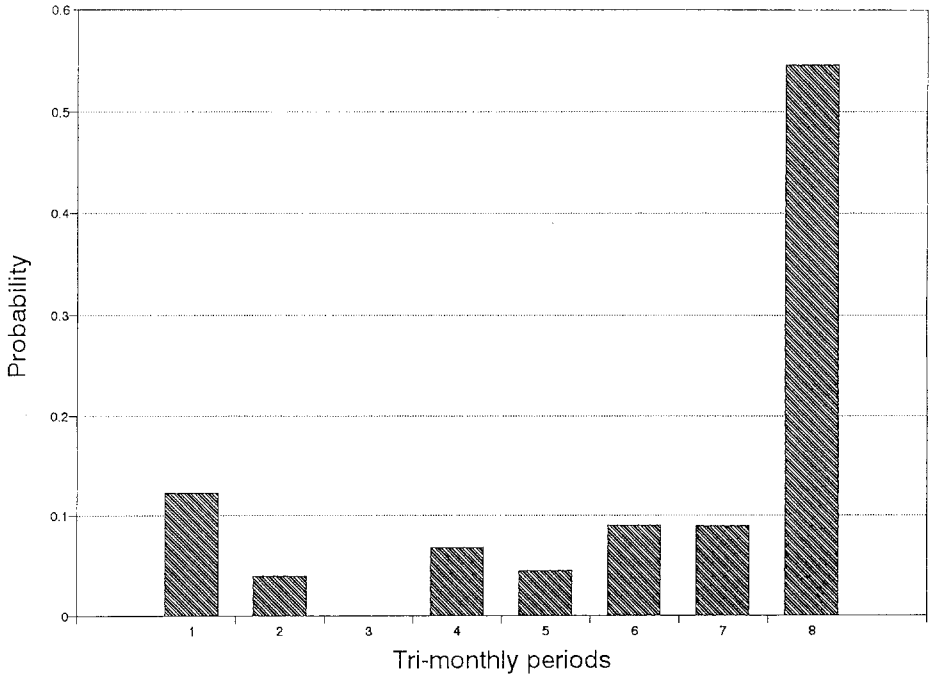

Fig. 5. Estimated time to change of the visit rate.

Forty of the 50 states increased their limits to 65 miles per hour, 38 of them in the spring or summer of 1987, and two in 1988.

The main goal of the analysis was to estimate the distribution of the time at which the traffic fatality rate changed. The assumption is that not all states would experience a change the instant the speed restriction was lifted.

Initially, an analysis for all 48 states with one or more fatalities was carried out even though the fatality rates differed greatly from state to state, violating the assumptions of Section 2. Not surprisingly, the estimated distribution placed virtually all its mass at either the beginning or the end of the time period under study. This phenomenon is due to the fact that the between state differences were most often much greater than the before-to-after differences within each state.

This problem was resolved by modifying the algorithm slightly, allowing λ_1 and λ_2 to vary from row to row but remain as fixed constants. Although the proof of consistency no longer holds in this case, further simulations seemed to show that the slightly modified algorithm performed well under conditions similar to those of this example. Since the rates λ_1 and λ_2 are estimated separately for each row, the modified algorithm always estimates $\hat{P}(N) = 0$. This is because the value of the likelihood contribution for each row cannot decrease when λ_1 and λ_2 are included in the model, that is, when $1 \leq \tau \leq 48$, compared to when the likelihood contains only λ_1, when $\tau = 49$. Thus the modified model is useful principally for estimation under the assumption that there is a change in each row. Information concerning the size of the change in each state is provided by the estimates of λ_1

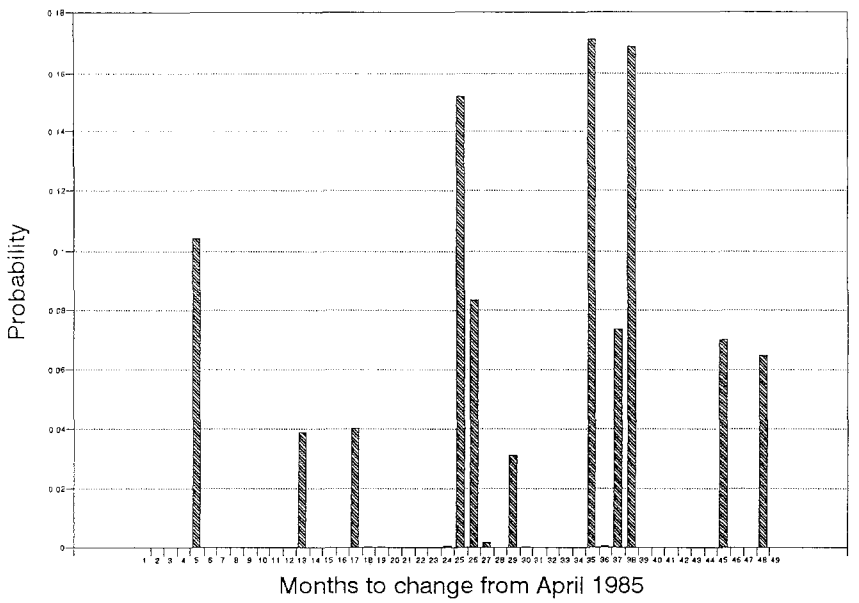

Fig. 6. Estimated time to change of the fatality rate.

and λ_2 for each row. It is presumed that the fatality rate does not affect P_T.

The resulting $\hat{P}_T$ appears in Fig. 6. It can be seen that 69% of the mass is concentrated around the months after the speed limit was lifted. In examining the estimated values for λ_1 and λ_2, 35 of the 48 states (73%) had $\hat{\lambda_1} > \hat{\lambda_2}$, indicating an increase in the fatality rates. Using the same stopping criterion as in the simulations, the algorithm converged in 114 iterations, taking approximately 50 seconds to run.

The data were not seasonally adjusted as this would have entailed a tedious state-by-state seasonal adjustment without further illuminating the method.

8. Concluding Remarks

The multi-path change-point problem is attacked on three fronts. By taking a missing data viewpoint, the EM algorithm is used to carry out the maximization. The same algorithm may be obtained by placing the estimation problem in the context of a more general mixture problem. Of course, there are other procedures for approximating maximum likelihood estimates. For a full discussion of these methods, including Newton's method and the conjugate gradient method, see Redner and Walker (1984). Alternatively, in order to prove the consistency of the maximum likelihood estimators, the work of Kiefer and Wolfowitz is invoked. They discuss estimation of parameters in the presence of a sequence of "random" parameters.

Careful simulations seem to indicate that all parameters are well estimated. Difficulty in estimating the change-point distribution may arise when the size of

the change is small and there are few data paths. Alternative methods for multi-path change-point problems are given in Joseph and Wolfson (1992).

Acknowledgements

The authors thank the referee for his/her meticulous reading of the manuscript and constructive suggestions, that led the authors to the implied generality of Theorem 5.1. The data for the urea formaldehyde foam insulation example were provided by Dr. John Hoey, of the Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada.

References

Billingsley, P. (1986). *Probability and Measure*, 2nd ed., Wiley, New York.

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

Cobb, G. W. (1978). The problem of the Nile: conditional solution to a change point problem, *Biometrika*, **62**, 243–251.

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.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables, *Biometrika*, **57**, 1–16.

Joseph, L. (1989). The multi-path change-point, Ph.D. Thesis, Department of Mathematics and Statistics, McGill University, Montreal.

Joseph, L. and Wolfson, D. B. (1992). Estimation in multi-path change-point problems, *Comm. Statist. Theory and Methods*, **21**, 897–913.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Math. Statist.*, **27**, 887–906.

L'Abbé, K. (1984). Health effects of urea-formaldehyde foam insulation, Master's Thesis, Department of Epidemiology and Biostatistics, McGill University, Montreal.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations, *Econometrica*, **16**, 1–32.

Peters, B. C. and Walker, H. F. (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions, *SIAM J. Appl. Math.*, **35**, 362–378.

Picard, D. (1985). Testing and estimating change-points in time series, *Adv. in Appl. Probab.*, **17**, 841–867.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions, *Ann. Statist.*, **9**, 225–228.

Redner, R. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev.*, **26**, 195–239.

Shaban, S. A. (1980). Change-point problem and two-phase regression: an annotated bibliography, *Internat. Statist. Rev.*, **48**, 83–93.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.*, **20**, 595–601.

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