Теорія Ймовір. та Матем. Статист.
Вип. 69, 2003

Theor. Probability and Math. Statist.
No. 69, 2004, Pages 103–113
S 0094-9000(05)00618-6
Article electronically published on February 8, 2005

THE LIMIT DISTRIBUTION OF DYNAMIC PROGRAMMING ESTIMATORS OF MULTIPLE CHANGE POINTS

UDC 519.21

R. E. MAĬBORODA AND O. V. SUGAKOVA

Abstract. We consider a problem of estimating multiple change points in the case where the distributions of observations between change points belong to a finite family of known distributions. We describe a dynamic programming procedure of the estimation and a method for improving estimators that generalizes the averaged likelihood method. The limit distributions of these estimators are given in terms of the argument of the minimum of random walks. We show that these distributions, for an appropriate set of parameters, coincide with those of the maximum likelihood estimators or averaged likelihood estimator for models with only one change point.

1. Introduction

There is extensive literature devoted to the change point analysis (see, for example, [1]). We consider the problem of estimating multiple change points of the distribution from independent observations in the case where the distributions between change points belong to a finite family of known distributions. Problems of this type appear often when analyzing geological data and also in algorithms of speech recognition. A dynamic programming procedure is developed in [2] to estimate the number of change points and their values (also see Section 2). It is shown in [3, 4] that the dynamic programming estimators are consistent for an appropriate set of parameters of the procedure. Upper bounds for the rate of convergence of these estimators are also obtained in [3, 4].

The limit distributions of estimators of change points are described in [5] (also see Section 3) for models with only one change point and with known distributions before and after the change point. It is shown in [5] that the limit probability of exact detection is maximal in the class of all asymptotically homogeneous estimators for maximum likelihood estimators. The limit mean square deviation is minimal for the so-called averaged likelihood estimators.

We describe the limit distributions of dynamic programming (DP) estimators for models with multiple change points (see Section 4). It turns out that for some set of parameters of the procedure, these distributions coincide with the distribution of the maximum likelihood estimator for the model with only one change point, obtained in [5]. This means that the DP procedure is able to achieve the best possible (limit) probability to detect the real values of change points irrespective of the number of change points and distributions between them.

In order to achieve the minimum of the limit mean square error we propose an additional procedure to improve the DP estimator. This procedure is described in Section 5. In Section 6 we discuss the questions on how one can apply the results of this paper and to what extent the results show that the DP estimators are indeed optimal.

2000 *Mathematics Subject Classification.* Primary 62G20; Secondary 93E10.

2. The setting of the problem

An observation $\Xi_N = \{\xi_1, \dots, \xi_N\}$ is a sequence of random elements of some measurable space $\mathcal{X}$. Let

$$\mathsf{P}\{\xi_j \in A\} = F_{h_j^0}(A),$$

where $F_h \in \mathfrak{F} = \{F_i, i = 1, \dots, K\}$ is a certain family of known distributions, and

$$\vec{h}_0 = \{h_j^0, j = 1, \dots, N\}$$

is an unknown nonrandom sequence of numbers (the trajectory of changes of distributions) such that $h_j^0 \in \{1, \dots, K\}$ and moreover $h_j^0 = \text{const}$ for $k_i \le j < k_{i+1}$, where $k_i = [\vartheta_i N]$ for $i = 0, \dots, R$ and some numbers $0 = \vartheta_0 < \vartheta_1 < \dots < \vartheta_R < \vartheta_{R+1} = 1$, $k_{R+1} = N + 1$.

Therefore there are R changes $k_1 < k_2 < \dots < k_R$ of the distribution in the sequence Ξ_N and moreover the random variables ξ_j, $j \in [k_i, k_{i+1})$, have the distribution $F_{h_{k_i}}$ of a given family $\mathfrak{F}$.

The numbers ϑ_i are called *points of change*. We express the limit behavior of k_i as $N \to \infty$ in terms of points of change. The exact values of ϑ_i, as well as exact values of k_i, are unknown. It is clear that we can construct an estimator $\hat{\vartheta}_{i,N} = \hat{k}_{i,N}/N$ for ϑ_i as long as an estimator $\hat{k}_{i,N}$ for k_i is constructed. When speaking about the consistency of estimation procedures for change points we usually mean that the estimators for points of change are consistent, that is,

$$\left|\hat{\vartheta}_{i,N} - \vartheta_i\right| \to 0$$

in probability as $N \to \infty$.

We deal with the problem of estimating R and k_i from the sample Ξ_N. Either the distributions $F_i \in \mathfrak{F}$ are known or only some of their parameters are available.

Denote by $\mathcal{H}$ the set of all possible trajectories of $\vec{h}$, that is, $\mathcal{H} = \{1, \dots, K\}^N$.

Note that the DP estimators approximate, in fact, a real trajectory $\vec{h}_0$. Consider functions $\phi\colon \mathcal{X} \times \{1, \dots, K\} \to \mathbb{R}$ and $\pi_N(g, l) = \pi_N \, \mathbb{1}\{g = l\}$ where π_N are some numbers. For an estimator for $\vec{h}_0$, we choose $\hat{h} = (\hat{h}_1, \dots, \hat{h}_N) \in \mathcal{H}$ such that the functional

$$J(h) = \sum_{j=1}^{N} \big(\pi_N(h_j, h_{j-1}) + \phi(\xi_j, h_j)\big) \tag{1}$$

attains its minimum at $\hat{h}$, that is,

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}} J(h). \tag{2}$$

Analogously the estimators for change points are

$$k_1(\hat{h}) = \hat{k}_{1,N} = \min\left\{l > 0\colon \hat{h}_l \ne \hat{h}_j, 1 \le j < l\right\}, \tag{3}$$

$$k_i(\hat{h}) = \hat{k}_{i,N} = \min\left\{l > \hat{k}_{i-1}\colon \hat{h}_l \ne \hat{h}_j, \hat{k}_{i-1} \le j < l\right\}, \quad \text{for } i = 2, 3, \dots, \tag{4}$$

and the estimator for the total number of changes is

$$R(\hat{h}) = \hat{R}_N = \min\left\{i\colon \hat{h}_l = \hat{h}_j \text{ for all } l, j,\ \hat{k}_{i,N} \le l, j \le N\right\}$$

(we assume that $\hat{k}_{0,N} = 0$ and $\hat{k}_{i,N} = N + 1$ if the set in (3) is empty). In other words, $\hat{k}_{i,N}$ are sequential change points in the trajectory $\hat{h}$, while $\hat{R}_N$ is the total number of these changes.

If $J(h)$ attains its minimum at several points h, then for the estimator $\hat{h}$ we choose the point of minimum of J for which $\sum_{i=1}^{R(h)} k_i(h)$ is minimal. (We prove later that under this convention the estimators are well defined with the probability approaching 1 as $N \to \infty$.) As the estimators for ϑ_i we take $\hat{\vartheta}_{i,N} = \hat{k}_{i,N}/N$.

By η_i we denote random elements in $\mathcal{X}$ with the distribution F_i.

It is shown in [3] that the estimators $\hat{\vartheta}_{i,N}$ are consistent if

$$\text{(5)} \qquad \text{(i)} \quad \mathsf{E}(\phi(\eta_i, l))^2 < \infty \quad \text{for all } i, l = 1, \dots, K;$$

$$\text{(6)} \qquad \text{(ii)} \quad \pi_N = CN^\beta \quad \text{for some } \frac{1}{2} < \beta < 1, 0 < C < \infty;$$

$$\text{(7)} \qquad \text{(iii)} \quad \Lambda := \inf_{i \neq k} (\mathsf{E}\,\phi(\eta_i, k) - \mathsf{E}\,\phi(\eta_i, i)) > 0.$$

Condition (iii) means that the expectation of $\phi(\xi_j, i)$ is minimal if i is the index of the true distribution of ξ_j. This property implies that $\hat{h}$ is a corresponding estimator for h^0.

Some extra conditions on the distributions of $\mathfrak{F}$ are needed in order that property (iii) be satisfied.

Example 1. Let the distributions F_i be known for all i. Put

$$\phi(x, i) = -\ln f_i(x) \quad \text{where } f_i := \frac{\partial F_i}{\partial \mu}(x).$$

Here μ is a measure such that all F_i are absolutely continuous with respect to μ. In this case, (7) holds in view of a well-known inequality

$$\int \ln f(x) g(x)\, \mu(dx) < \int \ln f(x) f(x)\, dx$$

for all $f \neq g$ almost everywhere with respect to the measure μ. Estimator (2) is called the generalized maximum likelihood estimator in this case.

Example 2. Let $\mathcal{X} = \mathbb{R}$, and let the distributions F_i be unknown but their expectations $m_k = \mathsf{E}\,\eta_k = \int x\, F_k(dx)$ be known, and suppose that $\mathsf{E}\,\eta_k^4 < \infty$. Putting $\phi(x,k) = (x - m_k)^2$, we obtain (7), since $\mathsf{E}(\eta - c)^2 > \mathsf{E}(\eta - \mathsf{E}\,\eta)^2$ for all $c \neq \mathsf{E}\,\eta$. Estimator (2) is called the generalized least square estimator.

Example 3. Let $\mathcal{X} = \mathbb{R}$, $K = 2$, and let

$$\operatorname{med}(F_1) > a, \qquad \operatorname{med}(F_2) < a$$

for some $a \in \mathbb{R}$ where $\operatorname{med}(F)$ is the median of the distribution F. Then the functions

$$\phi(x, 1) - \mathbb{1}\{x < a\}, \qquad \phi(x, 2) - \mathbb{1}\{x > a\}$$

satisfy condition (7). The corresponding DP estimator is called the median estimator.

The above three estimators differ in the amount of *a priori* information about the distributions needed to evaluate the estimator. Namely, the amount of information is maximal for the maximum likelihood estimator, moderate for the least squares estimator, and minimal for the median estimator. It is clear that the accuracy of the estimation also decreases with the amount of *a priori* information. To compare the quality of estimators for different cases when the sample is large (as $N \to \infty$) we consider the limit distributions of estimators $\hat{k}_{i,N}$, that is, the distributions of the random variables κ such that $\hat{k}_{i,N} - k_i \Rightarrow \kappa$. Before formulating the main results of this paper we recall some facts for the model with only one change point.

3. The model with only one change point

Assume that $R = 1$, $K = 2$, and $h_j^0 = \chi_1$ for $j < k := k_1$ and $h_j^0 = \chi_2$ for $j \geq k$. The problem is to estimate k. Put

$$J^0(s) = \sum_{j=1}^{s-1} \phi(\xi_j, \chi_1) + \sum_{j=s}^{N} \phi(\xi_j, \chi_2) \tag{8}$$

and $\hat{k}_N = \arg\min_s J^0(s)$. If the minimum is attained at several points s, then for $\hat{k}_N$ we choose the minimal of these points. It is clear that the estimator $\hat{k}_N$ coincides with $k_1(\hat{h})$ if $\hat{h}$ is defined by (2) where the minimum is taken with respect to the space of trajectories of the form $h = (\chi_1, \chi_1, \dots, \chi_1, \chi_2, \dots, \chi_2)$ instead of the space $\mathcal{H}$.

In what follows we assume that

$$\Psi_{i,l}(\lambda) = \mathsf{E}\exp(\lambda\phi(\eta_i, l)) < \infty \tag{9}$$

for all $\lambda \in \mathbb{R}$ and i, $l = 1, \dots, K$. Here $\Psi_{i,l}$ is the moment generating function of the random variable $\phi(\eta_i, l)$. Condition (9) is called the Cramér condition.

Denote by $\zeta_j = \zeta_j(\chi_1, \chi_2)$, $j \in \mathbb{Z}$, random variables such that

— the distribution of ζ_j coincides with the distribution of $-\phi(\eta_{\chi_1}, \chi_1) + \phi(\eta_{\chi_1}, \chi_2)$ for $j < 0$;

— $\zeta_j = 0$ for $j = 0$;

— the distribution of ζ_j coincides with the distribution of $\phi(\eta_{\chi_2}, \chi_1) - \phi(\eta_{\chi_2}, \chi_2)$ for $j > 0$.

Let

$$S_n(\chi_1, \chi_2) = S_n := \begin{cases} \sum_{j=1}^{n} \zeta_j & \text{if } n > 0; \\ \sum_{j=n}^{-1} \zeta_j & \text{if } n < 0 \end{cases} \tag{10}$$

for $n \in \mathbb{Z}$. Finally let

$$D = D(\chi_1, \chi_2) := \arg\min_{n \in \mathbb{Z}} S_n. \tag{11}$$

An estimator $\hat{k}_N$ for the true k constructed from Ξ_N is called asymptotically homogeneous if $p(\hat{k}_N) := \lim_{N\to\infty} \mathsf{P}_\vartheta(\hat{k}_N = k)$ does not depend on ϑ. (Here $k = [\vartheta N]$ and P_ϑ stands for the conditional probability given the true value of the change point in the sequence Ξ_N is ϑ.)

If F_{χ_1} and F_{χ_2} are known, then estimator (8) coincides with the maximum likelihood estimator for k if $\phi(x, i) = -\ln f_i(x)$ (see Example 1).

The following results are obtained in [5].

Proposition 1. *If* (9) *holds, then the estimator $\hat{k}_N$ defined by* (8) *is asymptotically homogeneous and $\hat{k}_N - k \Rightarrow D$.*

Note that this assertion is stated in Theorem 1 of [5] only for the maximum likelihood estimator; nevertheless the proof works in the general case, too.

Proposition 2 (Theorem 2 in [5])**.** *If $\tilde{k}_N$ is an arbitrary asymptotically homogeneous estimator and $\hat{k}_N^{MLE}$ is the maximum likelihood estimator for which* (9) *holds, then $p(\tilde{k}_N) \leq p(\hat{k}_N^{MLE})$.*

Therefore the probability to detect the true value of the change point is maximal as $N \to \infty$ among all asymptotically homogeneous estimators.

In contrast to regular estimation problems, the variance of the maximum likelihood estimator is not minimal for problems of the change point analysis. The variance of the so-called averaged likelihood estimator defined below is minimal for problems of the change point analysis.

Put

$$r(x,\chi_1,\chi_2) := \frac{f_{\chi_1}(x)}{f_{\chi_2}(x)}, \tag{12}$$

$$r_j = r_j(\chi_1,\chi_2) := r(\xi_j,\chi_1,\chi_2), \qquad R_n = R_n(\chi_1,\chi_2) = \prod_{j=1}^{n-1} r_j. \tag{13}$$

The averaged likelihood estimator for the change point k is defined by

$$\hat{k}_N^{ALE} = \sum_{n=1}^{N} nR_n \Big/ \sum_{n=1}^{N} R_n. \tag{14}$$

An estimator $\tilde{k}_N$ is called mean square asymptotically homogeneous if the limit

$$\sigma^2 = \sigma^2(\tilde{k}_N) := \lim_{N\to\infty} \mathsf{E}_\vartheta(\tilde{k}_N - k)^2 < \infty$$

exists and does not depend on the true value ϑ.

To determine the distribution of $\hat{k}_N^{ALE}$ we consider independent random variables ρ_j, $j \in \mathbb{Z}$, such that

— the distribution of ρ_j coincides with that of $(r(\eta_{\chi_1}))^{-1}$ for $j<0$;
— $\rho_j = 1$ for $j=0$;
— the distribution of ρ_j coincides with that of $r(\eta_{\chi_2})$ for $j>0$.

Put

$$E_{1,n} = \sum_{i=-n}^{-1} \prod_{j=i}^{1} \rho_j, \qquad E_{2,n} = \sum_{i=1}^{n} \prod_{j=1}^{i} \rho_j, \tag{15}$$

$$E'_{1,n} = -\sum_{i=-n}^{-1} i \prod_{j=i}^{1} \rho_j, \qquad E'_{2,n} = \sum_{i=1}^{n} i \prod_{j=1}^{i} \rho_j, \tag{16}$$

$$E_m = E_{m,\infty}, \qquad E'_m = E'_{m,\infty}, \qquad m=1,2, \tag{17}$$

$$H = H(\chi_1,\chi_2) = \frac{E'_2 - E'_1}{1+E_1+E_2}. \tag{18}$$

Proposition 3 (Theorem 4 in [5])**.** *If condition* (9) *holds, then the estimator* $\hat{k}_N^{ALE}$ *is mean square asymptotically homogeneous and* $\hat{k}_N^{ALE} - k \Rightarrow H$*. Moreover*

$$\sigma^2(\tilde{k}_N) \ge \sigma^2\big(\hat{k}_N^{ALE}\big)$$

for all mean square asymptotically homogeneous estimators $\tilde{k}_N$*.*

4. The asymptotic behavior of DP estimators

Now we turn to the case where the number of change points $R \ge 1$ is unknown. Denote by $\tilde{h}_i = h^0_{k_1}$ the index of the distribution of observations between the ith and $(i+1)$th change points. Let D_i, $i=1,\dots,R$, be independent random variables such that the distribution of D_i coincides with that of the random variable $D(\tilde{h}_{i-1},\tilde{h}_i)$ defined by (11), and let $\vec{D} = (D_1,\dots,D_R)$ and $\vec{k}_N = (\hat{k}_{1,N},\dots,\hat{k}_{R,N})$, where $\hat{k}_{i,N}$ are defined by (3) and (4), respectively, $\vec{k} = (k_1,\dots,k_R)$.

Theorem 4.1. *Let conditions* (6), (7), *and* (9) *hold. Then* $\vec{k}_N - \vec{k} \Rightarrow \vec{D}$ *as* $N\to\infty$*.*

Corollary 4.1. *Let all the assumptions of Theorem 4.1 hold. Then the distribution of the generalized maximum likelihood estimator considered in Example 1 coincides with that of the maximum likelihood estimator for the model with only one change point discussed in Proposition 1.*

Proof of Theorem 4.1. Put $\bar\vartheta = \min_{i=1,\dots,R+1} |\vartheta_i - \vartheta_{i-1}|$. According to a result in [3], there exists a constant $\mathcal{K}$ such that

$$\mathsf{P}\{A_N\} \ge 1 - \frac{4\sigma^2 \mathcal{K} N}{a_N^2} \tag{19}$$

where

$$A_N = \left\{\hat R = R,\ \hat h_{\hat k_{i,N}} = \tilde h_i,\ \text{and } |\hat k_{i,N} - k_i| \le 2(R+1)a_N/\Lambda \text{ for all } i = 0,1,\dots,R\right\}$$

and $\sigma^2 = \max_{i,l} \mathbf{D}\,\phi(\eta_i, l)$ if conditions (5) and (7) hold, $N > 3$, and

$$a_N < \frac{\pi_N}{R+3}, \qquad \pi_N < \frac{\Lambda\bar\vartheta N}{2(R+1)} \tag{20}$$

for some a_N. Note that (9) implies (5).

Put $a_N = N^\alpha$ for some $1/2 < \alpha < \beta$. Then (20) holds for large N in view of (6) and thus (19) implies that

$$\mathsf{P}(A_N) \ge 1 - 4\sigma\mathcal{K}N^{1-2\alpha} \to 1$$

as $N \to \infty$. Put

$$\mathcal{H}(b_N) = \left\{\vec h \in \mathcal{H}\colon R(h) = R, h_{k_i(h)} = \tilde h_i,\ \text{and } |k_i(h) - k_i| \le b_N \text{ for all } i = 0,\dots,R\right\}.$$

Thus $\mathcal{H}(b_N)$ is the set of all trajectories $\vec h$ that have the same number of change points as $\vec h^0$ does, and that may differ from $\vec h^0$ only for the numbers j such that $|j - k_i| \le b_N$ for some k_i.

Put $b_N = 2(R+1)N^\alpha/\Lambda$. In what follows we consider sufficiently large N for which the intervals $B_{i,N} = \{j\colon |j-k_i| \le b_N\}$ are disjoint (this can be achieved since $k_i = [\vartheta_i N]$).

Let

$$\check h_N = \operatorname*{arg\,min}_{\vec h \in \mathcal{H}(b_N)} J(\vec h).$$

Note that $\{\check h_N \ne \hat h_N\} \subseteq \bar A_N$, whence $\mathsf{P}\{\check h_N \ne \hat h_N\} \to 0$ as $N \to \infty$. Thus it remains to prove the assertion of the theorem only for $\check k_{i,N} = k_i(\check h_N)$ instead of $\hat k_{i,N} = k_i(\hat h_N)$.

Since all $h \in \mathcal{H}(b_N)$ have the same number of change points, the corresponding functional $J(h)$ contains the same number of terms π_N. Thus

$$\operatorname*{arg\,min}_{\vec h \in \mathcal{H}(b_N)} J(\vec h) = \operatorname*{arg\,min}_{\vec h \in \mathcal{H}(b_N)} \sum_{j=1}^N \phi(\xi_j, h_i).$$

It is convenient to express the functional to be minimized in terms of change points in $\vec h$. Let $s_0 = 1$, $s_{R+1} = N+1$, $s_i = k_i(\vec h)$, and $s = (s_1, \dots, s_R)$. Then

$$\sum_{j=1}^N \phi(\xi_j, h_i) = \sum_{l=0}^R \sum_{j=s_i}^{s_{i+1}-1} \phi(\xi_j, \tilde h_i) = J_1(s)$$

for $h \in \mathcal{H}(b_N)$, since the trajectories of $\mathcal{H}(b_N)$ have "correct" values of $\tilde h_i$ on the intervals between the ith and $(i+1)$th change points.

Thus the vector $\check k = (\check k_{1,N}, \dots, \check k_{R,N})$ can be written as

$$\check k = \operatorname*{arg\,min}_{s: s_i \in B_{i,N}} J_1(s).$$

Put $B^* = \{1, \dots, N\} \setminus \bigcup_{i=1}^{R} B_{i,N}$. Then

$$J_1(s) = J_2(s) + J_3(s)$$

where

$$J_2(s) := \sum_{l=1}^{R} J_2^{(l)}(s_l),$$
$$J_2^{(l)} := \sum_{k_l - b_N \le j < s_l} \phi(\xi_j, \tilde{h}_{l-1}) + \sum_{s_l \le j \le k_l + b_N} \phi(\xi_j, \tilde{h}_l),$$
$$J_3(s) := \sum_{j \in B^*} \phi\left(\xi_j, h_j^0\right).$$

It is clear that J_3 does not depend on s, whence

$$\check{k} = \underset{s\colon s_i \in B_{i,N}}{\arg\min} \sum_{l=1}^{R} J_2^{(l)}(s_l).$$

Each of the sums $J_2^{(l)}$ depends only on s_l. Thus one can minimize every term separately in these sums, that is,

$$\check{k}_{l,N} = \underset{s_l \in B_{l,N}}{\arg\max}\, J_2^{(l)}(s_l).$$

Since $J_2^{(l)}$ for different l depend on different random variables ξ_j and they are jointly independent, $J_2^{(l)}$ are also independent, hence the random variables $\check{k}_{l,N}$ are jointly independent too.

To complete the proof of the theorem, we need to show that $\check{k}_{l,N} - k_l \Rightarrow D_l$.

Subtracting $J_2^{(l)}([k_l + b_N])$ from $J_2^{(l)}(s)$ and changing the index of summation by $j \mapsto j - k$ we obtain

$$\check{k}_{l,N} - k_l = \arg\min J_4(u)$$

where

$$J_4(u) := \begin{cases} \sum_{j=1}^{u} \phi(\xi_{j+k_l}, \tilde{h}_{l-1}) - \phi(\xi_{j+k_l}, \tilde{h}_l), & \text{for } u > 0, \\ \sum_{j=u}^{1} -\phi(\xi_{j+k_l}, \tilde{h}_{l-1}) + \phi(\xi_{j+k_l}, \tilde{h}_l), & \text{for } u < 0. \end{cases}$$

It is clear that the distribution of $J_4(s)$ coincides with that of $S_u(\tilde{h}_{l-1}, \tilde{h}_l)$. Thus the distribution of $\check{k}_{l,N} - k_l$ coincides with that of

$$D_N^*(\tilde{h}_{l-1}, \tilde{h}_l) = \underset{|u| < b_N}{\arg\min}\, S_u(\tilde{h}_{l-1}, \tilde{h}_l).$$

It remains to check that $D_N^*(\tilde{h}_{l-1}, \tilde{h}_l) \to D(\tilde{h}_{l-1}, \tilde{h}_l)$ in probability as $N \to \infty$.

Since $S_0(\tilde{h}_{l-1}, \tilde{h}_l) = 0$, we have for all $b > 0$

$$\min_{|u| < b} S_u(\tilde{h}_{l-1}, \tilde{h}_l) \le 0.$$

Thus S_u cannot be minimal for $n \in \mathbb{Z}$ if $S_n(\tilde{h}_{l-1}, \tilde{h}_l) > 0$. Hence

$$\begin{aligned} E_N &:= \left\{ D_N^*(\tilde{h}_{l-1}, \tilde{h}_l) \ne D(\tilde{h}_{l-1}, \tilde{h}_l) \right\} \subseteq \left\{ \left| \underset{u \in \mathbb{Z}}{\arg\min}\, S_u(\tilde{h}_{l-1}, \tilde{h}_l) \right| \ge b_N \right\} \\ &\subseteq \left\{ \text{there exists } j\colon |j| \ge b_N, S_j(\tilde{h}_{l-1}, \tilde{h}_l) < 0 \right\} \end{aligned}$$

and

$$\mathsf{P}\{E_N\} \le \sum_{n\colon |n| \ge b_N} p_n \tag{21}$$

where $p_n := \mathsf{P}\{S_n(\tilde{h}_{l-1}, \tilde{h}_l) < 0\}$. Now we estimate p_n. It follows from (10) that

$$S_n := S_n(\tilde{h}_{l-1}, \tilde{h}_l) = \sum_{j=1}^{n} \zeta_j.$$

Condition (7) implies that $m := \mathsf{E}\,\zeta_j > 0$. Using Chernoff's inequality (see [6]) we get

$$\mathsf{P}\{S_n/|n| - m < x\} = \exp(-|n|\Psi^*(x))$$

for all $x < 0$ where $\Psi^*(x) = \sup_{\lambda \in \mathbb{R}}(\lambda x - \Psi(\lambda))$ is the Young–Fenchel transform of $\Psi(\lambda) = \ln \mathsf{E} \exp(\lambda(\zeta_1 - m))$. Note that $\Psi(\lambda) < \infty$ by (9). Now

$$p_n = \mathsf{P}\{S_n < 0\} \le \mathsf{P}\{S_n/|n| < m/2\} \le \mathsf{P}\{S_n/|n| - m < -m/2\} \le \exp(-|n|\Psi^*(-m/2)).$$

Since $\Psi^*(-m/2) > 0$ for any nondegenerate random variable ζ_1, we have

$$\sum_{n \ge b_n} p_n \le \sum_{n \ge b_n} \exp(-n\Psi^*(-m/2)) \to 0$$

as $N \to \infty$. Similarly $\sum_{n \le b_n} p_n \to 0$ and $\mathsf{P}\{E_N\} \to 0$ as $N \to \infty$ according to (21). The theorem is proved. □

5. The generalized averaged likelihood estimators

If the DP estimator $\hat{h}$ is constructed for the true trajectory $\vec{h}^0$, then one can construct the estimators $\hat{k}_{i,N} = k_{i,N}(\hat{h})$ for change points and $\check{h}_i = \hat{h}_{\hat{k}_{i,N}}$ for true indices of distributions $\tilde{h}_i$ on intervals between changes. One can improve estimators $\hat{k}_{i,N}$ by using the method of averaged estimators of the form (14). In doing so, we do not restrict the consideration to the functions $r(x, \chi_1, \chi_2)$ of the form (12), since the true densities $f_l(x)$ are unknown. We only assume that

$$m_1 := \mathsf{E} \ln r(\eta_{\chi_1}, \chi_1, \chi_2) > 0, \qquad m_2 := \mathsf{E} \ln r(\eta_{\chi_2}, \chi_1, \chi_2) < 0 \tag{22}$$

for all $1 \le \chi_1, \chi_2 \le K$.

Fix some $0 < \gamma < 1$. Put $k^-_{i,N} = [\hat{k}_{i,N} - \gamma(\hat{k}_{i,N} - \hat{k}_{i-1,N})]$ and

$$k^+_{i,N} = \left[\hat{k}_{i,N} + \gamma(\hat{k}_{i+1,N} - \hat{k}_{i,N})\right].$$

Consider

$$\bar{k}_{i,N} = \sum_{n=1}^{\hat{k}^+_{i,N} - \hat{k}^-_{i,N} + 1} n\hat{R}_n \Bigg/ \sum_{n=1}^{\hat{k}^+_{i,N} - \hat{k}^-_{i,N} + 1} \hat{R}_n + k^-_{i,N}$$

where

$$\hat{R}_n = \prod_{j = k^-_{i,N}}^{k^-_{i,N} + n} r(\xi_j, \check{h}_{i-1}, \check{h}_i).$$

Let $\bar{k}_N = (\bar{k}_{1,N}, \dots, \bar{k}_{R,N})$ and $H = (H_1, \dots, H_R)$ where $H_1, \dots, H_R$ are independent random variables such that the distribution of H_i coincides with that of $H(\tilde{h}_{i-1}, \tilde{h}_i)$ defined by (18).

Theorem 5.2. *Let all the assumptions of Theorem* 4.1 *hold. Assume also that* (22) *is satisfied. Then* $\bar{k}_N - k \Rightarrow H$.

Proof. Let $z_j = \ln \rho_j$ where ρ_j is defined in Section 3. It follows from (22) and the strong law of large numbers that

$$\frac{1}{i} \sum_{j=1}^{i} z_j \to m_2$$

almost surely as $i \to +\infty$. Thus

$$\frac{1}{i}\sum_{j=1}^{i} z_j < \frac{m_2}{2}$$

for all $i \geq i_0$ and some (random) i_0 (recall that $m_2 < 0$). Similarly

$$E_{2,N} = \sum_{i=1}^{n}\prod_{j=1}^{i} e^{z_j} \leq \sum_{i=1}^{i_0-1}\exp\biggl(\sum_{j=1}^{i} z_j\biggr) + \sum_{i=i_0}^{n}\exp(im/2),$$

whence the almost sure convergence of the series $E_2 = E_{2,\infty}$ follows and moreover

$$E_{2,n} \to E_2$$

almost surely as $n \to \infty$. We show similarly that $E_{1,n} \to E_1$ and $E'_{m,n} \to E'_m$, $m = 1, 2$, almost surely for all possible χ_1 and χ_2.

Let

$$\tilde{R}_n(a) := \prod_{j=a}^{a+n} r(\xi_j, \tilde{h}_{i-1}, \tilde{h}_i),$$

$$\tilde{k}_{i,N}(a, L) := \frac{\sum_{n=1}^{L} n\tilde{R}_n(a)}{\sum_{n=1}^{L}\tilde{R}_n(a)} + a.$$

As in the proof of Theorem 4.1 we obtain from (19) that

$$\mathsf{P}\left\{\hat{k}_{i,N} \neq \tilde{k}_{i,N}(k^-_{i,N}, [k^+_{i,N} - k^-_{i,N}] + 1)\right\} \to 0$$

as $N \to \infty$.

As before, it follows from (19) that $k_i - k^-_{i,N} \to \infty$ and $k^+_{i,N} - k_i \to \infty$ in probability and moreover

$$\mathsf{P}\left(\left\{k^-_{i,N} > k_{i-1}\right\} \cap \left\{k^+_{i,N} < k_{i+1}\right\}\right) \to 1.$$

Thus the desired assertion holds if

$$\tilde{k}_{i,N}(a_N, L_N) - k_i \Rightarrow H_i$$

as $N \to \infty$ for all nonrandom $a_N > 0$ and $L_N > 0$ such that

$$k_{i-1} < a_N < k_i, \qquad a_N + L_N < k_{i+1}, \qquad a_N + L_N - k_i \to \infty,$$

and $k_i - a_N \to \infty$. We omit the subscript N for both a and L for the sake of brevity.

Let $k^* := k_i - a$ and $r_j := r(\xi_{j+a}, \tilde{h}_{i-1}, \tilde{h}_i)$. It is easy to see that

$$\tilde{k}_{i,N}(a, L) - k_i = \frac{\sum_{l=1}^{L}(l - k^*)\prod_{j=1}^{l} r_j}{\sum_{l=1}^{L}\prod_{j=1}^{l} r_j}.$$

Dividing by $\prod_{j=a}^{k^*} r_j$ both the nominator and denominator on the right-hand side we get

$$\tilde{k}_{i,N}(a, L) - k_i = \frac{\sum_{l=1}^{k^*-1}(l - k^*)\prod_{j=l+1}^{k^*}(r_j)^{-1} + \sum_{l=k^*+1}^{L}(l - k^*)\prod_{j=k^*+1}^{l} r_j}{\sum_{l=1}^{k^*-1}\prod_{j=l+1}^{k^*}(r_j)^{-1} + \sum_{l=k^*+1}^{L}\prod_{j=k^*+1}^{l} r_j}.$$

Since the r_j are jointly independent and the random variable r_j for $j < k^*$ has the distribution $r(\eta_{\tilde{h}_{i-1}}, \tilde{h}_{i-1}, \tilde{h}_i)$, while its distribution for $j > k^*$ is $r(\eta_{\tilde{h}_i}, \tilde{h}_{i-1}, \tilde{h}_i)$, we get that $\tilde{k}_{i,N}(a, L) - k_i$ has the same distribution as the distribution of the random variable

$$\frac{-E'_{1,k^*-1}(\tilde{h}_{i-1}, \tilde{h}_i) + E'_{2,L-k^*}(\tilde{h}_{i-1}, \tilde{h}_i)}{E_{1,k^*-1}(\tilde{h}_{i-1}, \tilde{h}_i) + E_{2,L-k^*}(\tilde{h}_{i-1}, \tilde{h}_i) + 1}.$$

Since $E_{m,n}$ and $E'_{m,n}$ converge as $n \to \infty$, the proof of the theorem is complete. $\square$

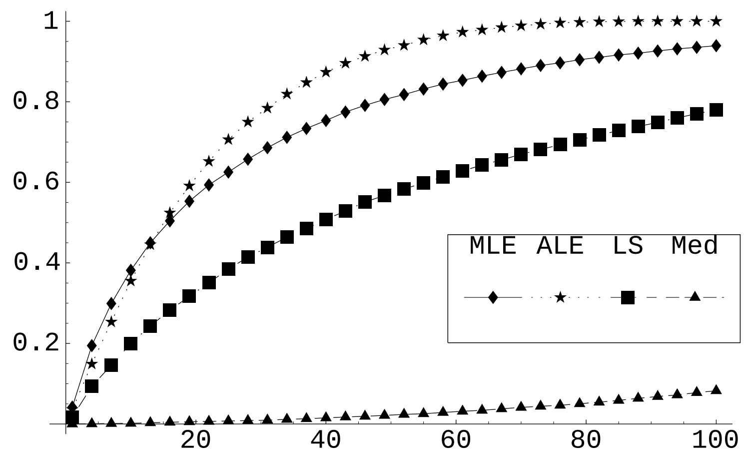

FIGURE 1. The p–Δ curves for some estimators of the change points.

6. Discussion

1. The asymptotic efficiency of one-dimensional parameters is usually characterized by the limit mean square error. From this point of view, the estimator $\bar{k}_{i,N}$ defined in Section 5 is asymptotically efficient. In regular cases, asymptotically efficient estimators are asymptotically optimal for practically all definitions of optimality (see [7]). This is not the case for the change point analysis.

It is natural to characterize the quality of an estimator $\hat{k}$ for k in terms of its reliability p and accuracy Δ:

$$p_{\hat{k}}(\Delta) = p := \mathsf{P}\left\{|\hat{k} - k| < \Delta\right\},$$

that is, the reliability of an estimator $\hat{k}$ is the probability that its deviation from the true value does not exceed a given accuracy. The larger $p_{\hat{k}}(\Delta)$ for a given accuracy Δ the better the estimator $\hat{k}$ for such an accuracy.

Our results allow us to construct the limit (as $N \to \infty$) curves $p = p(\Delta)$ (the so-called p–Δ curves) for DP estimators and for generalized averaged likelihood estimators.

The p–Δ curves in Figure 1 show four estimators:

— the generalized maximum likelihood estimator discussed in Example 1 (MLE);
— the generalized averaged likelihood estimator (ALE);
— the least squares estimator discussed in Example 2 (LS);
— the median estimator discussed in Example 3 (Med) for the parameter $a = 0.1$.

The distribution of data is normal with parameters $(0, 1)$ up to the change point and with parameters $(0.2, 1.2)$ after it. The curves are constructed from a simulation experiment with 10,000 samples.

As can be seen in Figure 1, if Δ is small, then the MLE has better reliability than the ALE, while the situation is opposite for large Δ. Therefore one cannot claim that a certain estimator of these two is better than the other one. The estimator LS is worse than the MLE and than the ALE; however, its reliability is still not bad. The reliability of the estimator Med is very low; it certainly cannot be used for $\Delta \leq 100$. However, if (as in the case discussed in Example 3) one cannot propose any estimate other than Med, then one should use it even though its accuracy is very low.

We should also mention that the p–Δ curves of these four estimators for other distributions can be absolutely different and therefore our conclusions about the estimators can also be different.

2. The asymptotic behavior of estimators obtained in this paper allows one to compare the limit distributions. However we do not study the rate of convergence to the limit distributions. For example, given different rates of increase of π_N in (6) we obtain DP estimators that are different in their quality; however, all of them have the same limit distribution described in Theorem 4.1. Therefore the asymptotic behavior obtained in this paper is too rough to solve the problem of the optimal choice of π_N.

3. An advantage of DP procedures is that they have linear complexity, that is, they require CNK elementary operations to evaluate the DP estimator (N is the sample size, K is the total number of distributions in $\mathfrak{F}$, and C is a constant). It is clear that the generalized averaged likelihood estimators defined in Section 5 also have the linear complexity; however, the constant C for them is greater.

Bibliography

1. L. Horváth, *The maximum likelihood method for testing changes in the parameters of normal observations*, Ann. Statist. **21** (1993), no. 2, 671–680. MR1232511 (94k:62028)
2. V. V. Mottl', I. B. Muchnik, and V. G. Yakovlev, *An optimal segmenting of experimental curves*, Avtomat. i Telemekh. **8** (1983), 84–95; English transl. in Automat. Remote Contr. (1984).
3. O. V. Sugakova, *A search for change points in a flow of independent observations*, Teor. Imovir. ta Matem. Statist. **55** (1996), 167–172; English transl. in Theor. Probab. and Math. Statist. **55** (1997), 181–186. MR1641589 (99g:62066)
4. R. E. Maĭboroda and O. V. Sugakova, *A fast algorithm for detecting multiple change points*, Teor. Imovir. ta Matem. Statist. **57** (1997), 103–108; English transl. in Theor. Probab. and Math. Statist. **57** (1998), 109–114. MR1806888 (2003b:62100)
5. A. A. Borovkov, *Asymptotical optimal solutions in the change point problem*, Teor. Veroyatnost. i Primenen. **43** (1998), no. 4, 625–654; English transl. in Theory Probab. Appl. **43** (1999), no. 4. MR1692429 (2001g:62044)
6. H. Chernoff, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Ann. Math. Statist. **23** (1952), no. 4, 493–507. MR0057518 (15:241c)
7. I. A. Ibragimov and R. Z. Khas'minskiĭ, *Statistical Estimation: Asymptotic Theory*, "Nauka", Moscow, 1979; English transl., Springer-Verlag, Berlin–Heidelberg–New York, 1981. MR0620321 (82g:62006)

Faculty for Mechanics and Mathematics, Kyiv National Taras Shevchenko University, Volodymyrs'ka Street 63, Kyiv, Ukraine
E-mail address: mre@mechmat.univ.kiev.ua

Faculty for Mechanics and Mathematics, Kyiv National Taras Shevchenko University, Volodymyrs'ka Street 63, Kyiv, Ukraine
E-mail address: sugak@univ.kiev.ua

Received 27/JAN/2003

Translated by S. KVASKO