The Annals of Statistics
1988, Vol. 16, No. 1, 188–197

NONPARAMETRIC CHANGE-POINT ESTIMATION[1]

By E. Carlstein

University of North Carolina at Chapel Hill

Consider a sequence of independent random variables $\{X_i: 1 \le i \le n\}$ having cdf F for $i \le \theta n$ and cdf G otherwise. A class of strongly consistent estimators for the change-point $\theta \in (0,1)$ is proposed. The estimators require no knowledge of the functional forms or parametric families of F and G. Furthermore, F and G need not differ in their means (or other measure of location). The only requirement is that F and G differ on a set of positive probability. The proof of consistency provides rates of convergence and bounds on the error probability for the estimators. The estimators are applied to two well-known data sets, in both cases yielding results in close agreement with previous parametric analyses. A simulation study is conducted, showing that the estimators perform well even when F and G share the same mean, variance and skewness.

1. Introduction. Let $X_1^n, \ldots, X_n^n$ be independent random variables with

$$X_1^n, \ldots, X_{[\theta n]}^n \text{ identically distributed with cdf } F,$$

$$X_{[\theta n]+1}^n, \ldots, X_n^n \text{ identically distributed with cdf } G,$$

where $[y]$ denotes the greatest integer not exceeding y. The parameter $\theta \in (0,1)$ is the change-point to be estimated. The body of literature addressing this problem is extensive, but most of the work is based upon at least one of the following assumptions:

1. F and G are known to belong to parametric families (e.g., normal, binomial) or are otherwise known in functional form.
2. F and G differ, in particular, in their levels (e.g., mean or median).

Hinkley (1970) and Hinkley and Hinkley (1970) use maximum likelihood to estimate θ in the situation where F and G are from the same parametric family. Hinkley (1972) generalizes this method to the case where F and G may be arbitrary known distributions, or alternatively where a sensible discriminant function (for discriminating between F and G) is known. Smith's (1975) Bayesian approach and Cobb's (1978) conditional solution also require assumptions of type 1. These authors generally suggest that any unknown parameters in F and G can be estimated from the sample, but nevertheless F and G must be specified as functions of those parameters.

At the other extreme, Darkhovshk (1976) presents a nonparametric estimator based on the Mann–Whitney statistic. Although his estimator makes no explicit use of the functional forms of F and G, his asymptotic results require

Received March 1985; revised December 1986.
[1]Research supported by NSF Grant DMS-84-00602.
AMS 1980 *subject classifications*. Primary 62G05; secondary 60F15.
Key words and phrases. Cramér–von Mises, distribution-free, Kolmogorov–Smirnov.

$\int_{-\infty}^{\infty} G(x)\,dF(x) \neq \frac{1}{2}$. This excludes cases where F and G are both symmetric and have a common median. Bhattacharyya and Johnson (1968) give a nonparametric test for the presence of a change-point, but again under the type 2 assumption that the variables after the change are stochastically larger than those before. [See Shaban (1980) for an annotated bibliography of change-point literature.]

In contrast to assumptions of types 1 and 2, the estimators studied here do not require any knowledge of F and G; virtually any salient difference between F and G will ensure detection of the change-point (asymptotically). Specifically, our basic assumption is:

$$(*)\qquad \text{The set}\quad \Lambda := \{x \in R\colon |F(x) - G(x)| > 0\}\quad \text{satisfies either}\quad \int_\Lambda dF(x) > 0 \quad\text{or}\quad \int_\Lambda dG(x) > 0.$$

Note that F and G may be discrete, continuous or mixed. The theoretical results for Darkhovshk's (1976) nonparametric estimator assumed F and G to be continuous. Unlike Cobb (1978), we do not require the supports of F and G to be identical; in fact the supports may be entirely unknown.

The intuition behind the proposed class of estimators is as follows. For a hypothetical (but not necessarily correct) change-point $t \in T_n := \{i/n\colon 1 \leq i \leq n-1\}$, consider the pre-$t$ empirical cdf ${}_th^n(x)$, which is constructed as if $X_1^n, \ldots, X_{nt}^n$ were identically distributed, and the post-t empirical cdf $h_t^n(x)$, which is constructed as if $X_{nt+1}^n, \ldots, X_n^n$ were identically distributed. That is,

$$\begin{aligned} {}_th^n(x) &:= \sum_{i=1}^{nt} I\{X_i^n \leq x\}\big/nt,\\ h_t^n(x) &:= \sum_{i=nt+1}^{n} I\{X_i^n \leq x\}\big/n(1-t). \end{aligned}$$

The former estimates the unknown mixture distribution

$${}_th(x) := I\{t \leq \theta\}F(x) + I\{t > \theta\}(\theta F(x) + (t-\theta)G(x))\big/t$$

and the latter similarly estimates

$$h_t(x) := I\{t \leq \theta\}((\theta - t)F(x) + (1-\theta)G(x))/(1-t) + I\{t > \theta\}G(x).$$

The difference between these two unknown distributions is measured by the entities

$$\begin{aligned} \delta_{ni}^t &:= |{}_th(X_i^n) - h_t(X_i^n)|\\ &= (I\{t \leq \theta\}(1-\theta)/(1-t) + I\{t > \theta\}\theta/t)\delta_{ni}^\theta, \qquad 1 \leq i \leq n. \end{aligned}$$

Now, combining these n differences via any homogeneous norming function $S_n\colon R^n \to R$ [i.e., $S_n(cy_1, \ldots, cy_n) = cS_n(y_1, \ldots, y_n) \geq 0$, whenever $c \geq 0$ and $y_i \geq 0\ \forall\ i$] we obtain

$$\begin{aligned} \Delta_n(t) &:= t^{1/2}(1-t)^{1/2}S_n(\delta_{n1}^t, \ldots, \delta_{nn}^t)\\ &= \rho(t)S_n(\delta_{n1}^\theta, \ldots, \delta_{nn}^\theta), \end{aligned}$$

where $\rho(t) := I\{t \le \theta\}(1-\theta)t^{1/2}/(1-t)^{1/2} + I\{t > \theta\}\theta(1-t)^{1/2}/t^{1/2}$. Note that $\Delta_n(t)$ attains its maximum [over $t \in (0,1)$] at $t = \theta$. Thus a reasonable estimator for θ is the value of $t \in T_n$ which maximizes the corresponding sample criterion function

$$D_n(t) := t^{1/2}(1-t)^{1/2} S_n\big(d_{n1}^t, \dots, d_{nn}^t\big),$$

where

$$d_{ni}^t := |{}_t h^n(X_i^n) - h_t^n(X_i^n)|, \qquad 1 \le i \le n.$$

In Section 2 the estimators are discussed in more detail and their asymptotic properties are presented. The results include strong consistency, with rates of convergence and bounds on the error probability. Proof of these results is deferred to Section 4. Section 3 investigates the finite-sample behavior of the estimators: First the estimators are calculated on Cobb's (1978) Nile data and on the Lindisfarne scribes data [see Smith (1980)]. In both examples this nonparametric analysis produces results which are nearly identical to the results of the earlier parametric analyses. Then the estimators are tested (via simulation) in a situation where no other change-point estimator can be used: F and G are both unknown and are of different parametric families, but they are both symmetric and share the same mean and variance. Here again the performance of the estimators is quite reasonable.

2. Properties of the estimators. The change-point estimator described in Section 1 is formally defined as

$$\theta_n \in T_n \quad \text{for which} \quad D_n(\theta_n) = \max_{t \in T_n} D_n(t).$$

Notice that there are only $n-1$ distinct values of $D_n(\cdot)$ to be compared, so θ_n always exists and is easily calculated.

Certain constraints must be imposed upon the choice of a norm S_n. The following conditions are intuitively reasonable, convenient for proving theoretical results about θ_n and easy to check in practice. For $n \ge 1$, denote $\mathbf{y}_n := (y_1, \dots, y_n)$, $\mathbf{0}_n := (0, \dots, 0)$ and $\mathbf{1}_n := (1, \dots, 1)$. Write $\mathbf{y}_n^{(1)} \ge \mathbf{y}_n^{(2)}$ for the condition $y_i^{(1)} \ge y_i^{(2)}\ \forall\ i \in \{1, \dots, n\}$.

DEFINITION. A function S_n: $R^n \to R$ is a *mean-dominant norm* iff:

1. (Symmetry) S_n is symmetric in its n arguments.
2. (Homogeneity) $S_n(c\mathbf{y}_n) = cS_n(\mathbf{y}_n)$ whenever $c \ge 0$ and $\mathbf{y}_n \ge \mathbf{0}_n$.
3. (Triangle inequality) $S_n(\mathbf{y}_n^{(1)} + \mathbf{y}_n^{(2)}) \le S_n(\mathbf{y}_n^{(1)}) + S_n(\mathbf{y}_n^{(2)})$ whenever $\mathbf{y}_n^{(1)} \ge \mathbf{0}_n$ and $\mathbf{y}_n^{(2)} \ge \mathbf{0}_n$.
4. (Identity) $S_n(\mathbf{1}_n) = 1$.
5. (Monotonicity) $S_n(\mathbf{y}_n^{(1)}) \ge S_n(\mathbf{y}_n^{(2)})$ whenever $\mathbf{y}_n^{(1)} \ge \mathbf{y}_n^{(2)} \ge \mathbf{0}_n$.
6. (Mean dominance) $S_n(\mathbf{y}_n) \ge \sum_{i=1}^n y_i/n$ whenever $\mathbf{y}_n \ge \mathbf{0}_n$.

Some natural examples of mean-dominant norms are $S_n^{(1)}(\mathbf{y}_n) := \sum_{i=1}^n y_i/n$,

$S_n^{(2)}(\mathbf{y}_n) := (\sum_{i=1}^n y_i^2/n)^{1/2}$ and $S_n^{(3)}(\mathbf{y}_n) := \sup_{1\le i\le n} y_i$. Let $D_n^{(i)}(t)$ and $\theta_n^{(i)}$ denote the criterion function and estimator, respectively, corresponding to norm $S_n^{(i)}$, $1 \le i \le 3$. $D_n^{(2)}(t)$ is a Cramér–von Mises distance between the cdfs ${}_t h^n(\cdot)$ and $h_t^n(\cdot)$, $D_n^{(3)}$ is similarly a Kolmogorov–Smirnov distance.

Since θ_n is being proposed as a nonparametric estimator, it would be desirable for θ_n to be invariant under strict monotone (increasing or decreasing) transformations of the data $X_1, \dots, X_n$. Since $D_n^{(3)}(t)$ is unchanged by strict monotone transformations of the data, the particular estimator $\theta_n^{(3)}$ has the desired invariance property. For an arbitrary mean-dominant norm S_n, the corresponding criterion function $D_n(t)$ and estimator θ_n are invariant *if* the observations $X_1, \dots, X_n$ are all distinct. If, on the other hand, the data contain ties, then $D_n(t)$ is not in general invariant. This is due to the arbitrary use of *lower* cdfs [e.g., $F(x) := P\{X \le x\}$] rather than *upper* cdfs [e.g., $\tilde{\bar{F}}(x) := P\{X \ge x\}$] in our estimation procedure. The following modification will always yield an invariant estimator. Calculate $\tilde{d}_{ni}^t := |{}_t\tilde{h}^n(X_i^n) - \tilde{h}_t^n(X_i^n)|$ for each $i \in \{1, \dots, n\}$ and $t \in T_n$, where ${}_t\tilde{h}^n(x) := \sum_{i=1}^{nt} I\{X_i^n \ge x\}/nt$ and $\tilde{h}_t^n(x) := \sum_{i=nt+1}^{n} I\{X_i^n \ge x\}/n(1-t)$. Using analogous logic to that in Section 1, the maximizer $\tilde{\theta}_n \in T_n$ of $\tilde{D}_n(t) := t^{1/2}(1-t)^{1/2} S_n(\tilde{d}_{n1}^t, \dots, \tilde{d}_{nn}^t)$ is also a reasonable estimator of θ. Define $\hat{\theta}_n := (\theta_n + \tilde{\theta}_n)/2$ as our modified estimator. Now, both $D_n(t)$ and $\tilde{D}_n(t)$ are invariant under strictly increasing transformations of the data, so θ_n, $\tilde{\theta}_n$ and $\hat{\theta}_n$ are invariant under such transformations. When a strictly decreasing transformation is applied to the data, $D_n(t)$ calculated from the transformed data is equal to $\tilde{D}_n(t)$ calculated from the original data; likewise $\tilde{D}_n(t)$ calculated from the transformed data is equal to $D_n(t)$ calculated from the original data. Thus $\hat{\theta}_n$ is invariant under *all* strict monotone transformations of the data, regardless of the choice of S_n and regardless of the presence of ties in the data.

The theoretical properties of θ_n include strong consistency and an exponential bound on the error probability.

THEOREM 1. *Let $\{S_n\colon n \ge 1\}$ be mean-dominant norms and assume $(*)$ holds. Let $\delta \in [0, \frac{1}{2})$ be arbitrary but fixed. Then*

$$|\theta_n - \theta| n^{\delta} \to 0 \quad \textit{almost surely as } n \to \infty.$$

THEOREM 2. *Let $\{S_n\colon n \ge 1\}$ be mean-dominant norms and assume $(*)$ holds. Then, for any $\varepsilon > 0$,*

$$P\{|\theta_n - \theta| > \varepsilon\} \le c_1 n \exp\{-c_2 \varepsilon^2 n\} \qquad \forall\, n \ge n(\varepsilon),$$

where $c_1 > 0$ and $c_2 > 0$ are constants.

Proof of these results is deferred to Section 4. Analogous results hold for $\tilde{\theta}_n$.

3. Applications.

The Nile data. Cobb (1978) reports the annual volume of discharge from the Nile River for each year from 1871–1970. His analyses assume that the observations are independent normal variables with common variance for the whole

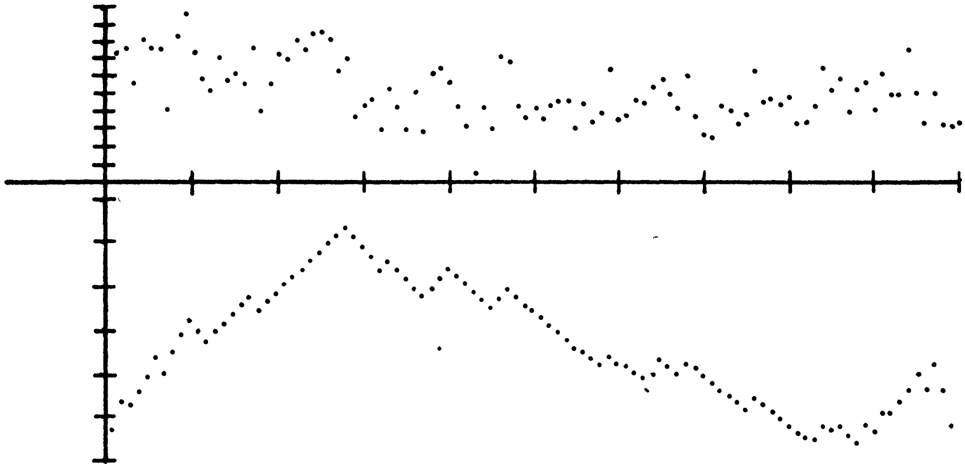

FIG. 1. *The Nile data. i is the year* (1871 *to* 1970), X_i *is the annual volume of discharge* (10^{10} m^3).

series. The results he obtains clearly indicate 1898 as the most likely change-point; he cites independent meteorological evidence that this change is real. Figure 1 shows the data X_i along with the criterion function $D_n^{(3)}(t)$; the corresponding nonparametric estimate is $\theta_n^{(3)} = 0.28$ (i.e., 1898). Since there are ties in the data, $D_n^{(1)}(t) \neq \tilde{D}_n^{(1)}(t)$ and $D_n^{(2)}(t) \neq \tilde{D}_n^{(2)}(t)$. But the maximizers $\theta_n^{(1)}$, $\tilde{\theta}_n^{(1)}$, $\theta_n^{(2)}$ and $\tilde{\theta}_n^{(2)}$ are nevertheless all equal to 0.28.

The Lindisfarne scribes data. The Lindisfarne text [as studied by Smith (1980)] divides into 13 sections. It is assumed that only one scribe was involved in the writing of any one section, and that sections written by any one scribe are consecutive. It is also assumed that a scribe may be characterized by his propensity to use one of two possible grammatical variants: either the "s" or "δ" ending in the present indicative third person singular. Let m_j denote the total number of relevant words in the jth section, and let Y_j denote the number of times that the "s" ending was used in those words. Smith (1980) assumes that the Y_j are independent binomial variables with common parameter p between change-points and he uses independent beta prior distributions on the p's. His analysis (which entertains the possibility of multiple change-points) arrives at a model with changes of scribe after Section 4 and again after Section 5.

Let $n := \sum_{j=1}^{13} m_j$ represent the total number of relevant words in the entire text and let X_i $(1 \leq i \leq n)$ be an indicator variable for the "δ" ending in the ith word. Our analysis assumes that the X_i are independent Bernoulli r.v.s with common parameter p_1 for all $i \leq [\theta n]$ and common parameter p_2 for all $i \geq [\theta n] + 1$ (p_1 and p_2 unknown). In this scenario, $D_n(t)$ [and $\tilde{D}_n(t)$] are proportional to

$$C_n(t) := t^{1/2}(1-t)^{1/2}\left|\sum_{i=1}^{nt} X_i/t - \sum_{i=nt+1}^{n} X_i/(1-t)\right|$$

TABLE 1
The Lindisfarne scribes data[a]

k	1	2	3	4	5	6	7	8	9	10	11	12	13
n_k	21	57	101	131	183	228	276	333	381	403	423	444	464
$\sum_{i=1}^{n_k} X_i$	9	19	32	38	62	73	82	93	100	103	106	110	114
$C_n(n_k/n)$	18.5	15.2	17.4	12.9	34.9	34.0	28.9	24.8	16.7	11.8	7.3	4.5	—

[a] k is the section of the text, n_k is the number of relevant words in sections 1 through k, X_i is the indicator of "ð" ending in ith word.

for all choices of mean-dominant norm S_n. Since we only consider potential change-points at the *ends* of sections, the criterion function $C_n(t)$ is computed only for values of t in the restricted set $\{\sum_{j=1}^{k} m_j/n\colon 1 \le k \le 12\} \subset T_n$. Table 1 shows the data and the function $C_n(t)$, which is maximized at $\theta_n = 183/464$ (i.e., at the end of Section 5).

Simulation study. When the functional forms of F and G are unknown to the statistician, but both distributions are symmetric with the same mean, then no other change-point estimator is appropriate. Such is the case in this example:

$$F \text{ is the distribution with density } f(x) = 0.697128x^2 I\{|x| < 1.291\};$$

$$G \text{ is the } N(0,1) \text{ distribution}.$$

Actually, F and G also share the same variance in this situation, making it even more difficult for the user to choose an estimator that discriminates between them. Table 2 presents simulation results for $\theta_n^{(1)}$, $\theta_n^{(2)}$ and $\theta_n^{(3)}$, based on sample sizes $n = 50$, $n = 100$ and $n = 200$. For the moderately large sample size $n = 200$, all three estimators perform well (and in fact no one estimator appears to be

TABLE 2
Simulation study. True value of $\theta = 0.4$. The entries in each row are empirical estimates based on 1000 realizations of $\theta_n^{(i)}$. The standard deviation of each estimated $E\{\theta_n^{(i)}\}$ is less than 0.01.

n	i	$E\{\theta_n^{(i)}\}$	$E\{\lvert\theta_n^{(i)} - \theta\rvert\}$
50	1	0.443	0.257
	2	0.418	0.235
	3	0.400	0.179
100	1	0.420	0.201
	2	0.401	0.178
	3	0.392	0.144
200	1	0.404	0.0971
	2	0.391	0.0967
	3	0.390	0.0957

clearly superior to the others). For the smaller sample sizes, the choice of S_n has a more pronounced effect: $\theta_n^{(1)}$ is clearly the worst, while $\theta_n^{(3)}$ is arguably the best. This empirical evidence in favor of $\theta_n^{(3)}$, together with its invariance property (discussed in Section 2) and its intuitive appeal (being based upon a Kolmogorov–Smirnov distance), may convince the practitioner that $\theta_n^{(3)}$ is the preferable choice for nonparametric change-point estimation.

4. Proofs. Assume that $\{S_n\}$ are mean-dominant and that (*) holds. Let $\delta \in [0, \frac{1}{2})$ be fixed. We shall show that, for any $\varepsilon > 0$,

$$P\{|\theta_n - \theta| n^{\delta} > \varepsilon\} \le c_1 n \exp\{-c_2 \varepsilon^2 n^{1-2\delta}\} \qquad \forall\, n \ge n(\varepsilon, \delta),$$

where $c_1 > 0$ and $c_2 > 0$ are constants. Theorem 1 then follows by applying the Borel–Cantelli lemma. Theorem 2 is obtained by setting $\delta = 0$.

Throughout this section, the entities K_i will denote finite positive constants. Define $\overline{T}_n := T_n \cup \{0/n, n/n\}$ and $\bar{\theta} := \theta^{1/2}(1-\theta)^{1/2}$.

LEMMA 1. *Let $Y_1^n, \dots, Y_n^n$ be iid with cdf Q; let r, s, l and m each be elements of $\overline{T}_n$ and satisfy $m \le s < r \le l$. Define*

$$Q_n(x; r, s, l, m) := \sum_{i=sn+1}^{rn} I\{Y_i^n \le x\}/(l-m)n,$$

$$q_n(x; r, s, l, m) := |Q_n(x; r, s, l, m) - (r-s)Q(x)/(l-m)|,$$

$$\hat{q}_n(r, s) := \sup_{\{m \in \overline{T}_n,\, l \in \overline{T}_n:\ m \le s \text{ and } l \ge r\}} \ \sup_{x \in R} q_n(x; r, s, l, m).$$

Then

$$P\{\hat{q}_n(r, s) n^{\delta} > \varepsilon\} \le K_1 \exp\{-K_2 \varepsilon^2 (r-s) n^{1-2\delta}\} \qquad \forall\, n.$$

PROOF. Since $q_n(x; r, s, l, m) \le q_n(x; r, s, r, s)$, we have $\hat{q}_n(r, s) \le \sup_{x \in R} q_n(x; r, s, r, s)$. Now apply Lemma 2 of Dvoretzky, Kiefer and Wolfowitz (1956) in order to bound $P\{\sup_{x \in R} q_n(x; r, s, r, s) > \varepsilon n^{-\delta}\}$; see also the discussion after their Theorem 3. □

LEMMA 2.

$$P\Big\{\sup_{t \in T_n} |D_n(t) - \Delta_n(t)| n^{\delta} > \varepsilon\Big\} \le K_3 n \exp\{-K_4 \varepsilon^2 n^{1-2\delta}\} \qquad \forall\, n \ge N(\varepsilon, \delta).$$

PROOF. Denote ${}_tH^{ni} := |{}_th^n(X_i^n) - {}_th(X_i^n)|$, $H_t^{ni} := |h_t^n(X_i^n) - h_t(X_i^n)|$ and $e_{ni}^t := {}_tH^{ni} + H_t^{ni}$. Since $d_{ni}^t \le \delta_{ni}^t + e_{ni}^t$, we have

$$D_n(t) - \Delta_n(t) \le t^{1/2}(1-t)^{1/2} S_n(e_{n1}^t, \dots, e_{nn}^t)$$

by virtue of S_n's properties (5) and (3). The same bound applies to $\Delta_n(t) - D_n(t)$, yielding [by (3)]

$$|D_n(t) - \Delta_n(t)| \le t^{1/2}(1-t)^{1/2}\big(S_n({}_tH^{n1},\dots,{}_tH^{nn}) + S_n(H_t^{n1},\dots,H_t^{nn})\big).$$

Now, for $t \in T_n$,

$$\begin{aligned}
{}_tH^{ni} &\le I\{t \le \theta\}|F_n(X_i^n; t,0,t,0) - F(X_i^n)| \\
&\quad + I\{t > \theta\}\big(|F_n(X_i^n; [\theta n]/n, 0, t, 0) - \theta F(X_i^n)/t| \\
&\qquad + |G_n(X_i^n; t, [\theta n]/n, t, 0) - (t-\theta)G(X_i^n)/t|\big) \\
&\le I\{t \le \theta\} f_n(X_i^n; t,0,t,0) \\
&\quad + I\{t > \theta\}\big(f_n(X_i^n; [\theta n]/n, 0, t, 0) + |[\theta n]/n - \theta| F(X_i^n)/t \\
&\qquad + g_n(X_i^n; t, [\theta n]/n, t, 0) + |[\theta n]/n - \theta| G(X_i^n)/t\big) \\
&\le I\{t \le \theta\}\hat{f}_n(t,0) \\
&\quad + I\{t > \theta\}\big(\hat{f}_n([\theta n]/n, 0) + \hat{g}_n(t, [\theta n]/n)(t - [\theta n]/n)/t + 2/n\theta\big) \\
&:= {}_tH^{n*}.
\end{aligned}$$

Here we use the notation of Lemma 1 (e.g., F_n, F, f_n and $\hat{f}_n$ play the role of Q_n, Q, q_n and $\hat{q}_n$ when the corresponding X_i^n are iid with cdf F). A similar bound (call it H_t^{n*}) applies to H_t^{ni}. Since these bounds on ${}_tH^{ni}$ and H_t^{ni} do not depend on i, properties (5), (2) and (4) yield $|D_n(t) - \Delta_n(t)| \le t^{1/2}(1-t)^{1/2}({}_tH^{n*} + H_t^{n*})$. Next we shall bound $P\{\sup_{t\in T_n} t^{1/2}(1-t)^{1/2}{}_tH^{n*} > \varepsilon/2n^{\delta}\}$; an analogous argument applies to H_t^{n*}. Since T_n is a finite set, the preceding probability is bounded by

$$\sum_{t\in T_n,\, t\le\theta} P\{t^{1/2}\hat{f}_n(t,0) > \varepsilon/2n^{\delta}\} + \sum_{t\in T_n,\, t>\theta} \big(P\{\hat{f}_n([\theta n]/n,0) > \varepsilon/6n^{\delta}\}$$
$$+ P\{\hat{g}_n(t,[\theta n]/n)(t - [\theta n]/n)/t > \varepsilon/6n^{\delta}\} + P\{2/n\theta > \varepsilon/6n^{\delta}\}\big).$$

Using Lemma 1, the probability in the first summation and the first probability in the second summation are each bounded by $K_1\exp\{-K_2'\varepsilon^2 n^{1-2\delta}\}$. Again using Lemma 1 and the fact that $t > \theta$, the second probability in the second summation is bounded by $K_1\exp\{-K_2''\varepsilon^2 n^{1-2\delta}\}$. The third probability in the second summation is zero for n sufficiently large. Finally, each summation is taken over order-n values of t. □

The maximizer of $\Delta_n(\cdot)$ over the set T_n is

$$\begin{aligned}
t_n := {}& I\{\rho([\theta n]/n) \ge \rho(([\theta n]+1)/n)\}[\theta n]/n \\
&+ I\{\rho([\theta n]/n) < \rho(([\theta n]+1)/n)\}([\theta n]+1)/n.
\end{aligned}$$

LEMMA 3.

$$P\{|\Delta_n(\theta_n) - \Delta_n(\theta)|n^{\delta} > \varepsilon\} \le K_5 n \exp\{-K_6\varepsilon^2 n^{1-2\delta}\} \qquad \forall\, n \ge N'(\varepsilon,\delta).$$

PROOF.

$$|\Delta_n(\theta_n) - \Delta_n(\theta)| \le |\Delta_n(\theta_n) - D_n(\theta_n)| + |D_n(\theta_n) - \Delta_n(t_n)| + |\Delta_n(t_n) - \Delta_n(\theta)|.$$

Note that either $D_n(\theta_n) \geq \Delta_n(t_n) \geq \Delta_n(\theta_n)$ or $\Delta_n(t_n) \geq D_n(\theta_n) \geq D_n(t_n)$; in either case the second term on the r.h.s. of the preceding inequality is bounded by $\sup_{t \in T_n}|D_n(t) - \Delta_n(t)|$. The same bound applies to the first term. Hence both of these terms are handled by Lemma 2.

The third term equals

$$\begin{aligned}(\rho(\theta) - \rho(t_n))S_n(\delta^\theta_{n1},\dots,\delta^\theta_{nn}) &\leq \rho(\theta) - \rho(t_n) \\ &= \bar{\theta}\big(I\{t_n \leq \theta\}(1 - b_n^{1/2}) + I\{t_n > \theta\}(1 - b_n^{-1/2})\big) \\ &\leq \bar{\theta}\big(I\{t_n \leq \theta\}(1 - b_n) + I\{t_n > \theta\}(1 - b_n^{-1})\big),\end{aligned}$$

where $b_n := t_n(1-\theta)/\theta(1-t_n)$ and $\delta^\theta_{ni} \leq 1$. Note that t_n, being the maximizer of $\rho(\cdot)$ over T_n, is nonrandom. Hence it suffices to observe that

$$\begin{aligned}&\bar{\theta}\big(I\{t_n \leq \theta\}(\theta - t_n)/\theta(1-t_n) + I\{t_n > \theta\}(t_n - \theta)/t_n(1-\theta)\big) \\ &\quad \leq \bar{\theta}^{-1}\big(I\{t_n \leq \theta\}(\theta - t_n) + I\{t_n > \theta\}(t_n - \theta)\big) \leq K_0 n^{-1} \leq \varepsilon n^{-\delta}\end{aligned}$$

for n sufficiently large. □

Denote $\mu_F := \int_R |F(x) - G(x)|\, dF(x)$, $\mu_G := \int_R |F(x) - G(x)|\, dG(x)$, $\mu := \theta\mu_F + (1-\theta)\mu_G$ and $c := \mu/2 > 0$ [by (*)]. Note that $\delta_n := S_n^{(1)}(\delta^\theta_{n1},\dots,\delta^\theta_{nn})$ may be written as $\underline{\delta}_n[\theta n]/n + \bar{\delta}_n(n - [\theta n])/n$, where $\underline{\delta}_n := \sum_{i=1}^{[\theta n]} \delta^\theta_{ni}/[\theta n]$ and $\bar{\delta}_n := \sum_{i=[\theta n]+1}^{n} \delta^\theta_{ni}/(n - [\theta n])$.

Lemma 4.

$$P\{|\theta_n - \theta| n^\delta > \varepsilon\} \leq K_7 n \exp\{-K_8 \varepsilon^2 n^{1-2\delta}\} \qquad \forall\, n \geq N''(\varepsilon, \delta).$$

Proof. Consider $t \in (0,1)$ and $\nu > 0$. Denote $b_t := t(1-\theta)/\theta(1-t)$. Recalling property (6),

$$\begin{aligned}|t - \theta| > \nu \Rightarrow |\Delta_n(t) - \Delta_n(\theta)| &= \bar{\theta}\big(I\{t \leq \theta\}(1 - b_t^{1/2}) + I\{t > \theta\}(1 - b_t^{-1/2})\big) \\ &\quad \times S_n(\delta^\theta_{n1},\dots,\delta^\theta_{nn}) \\ &\geq \bar{\theta}\big(I\{t \leq \theta\}(1 - b_t)/2 + I\{t > \theta\}(1 - b_t^{-1})/2\big)\delta_n \\ &\geq \tfrac{1}{2}\bar{\theta}\big(I\{t \leq \theta\}(\theta - t) + I\{t > \theta\}(t - \theta)\big)\delta_n \geq K_0' \nu\, \delta_n,\end{aligned}$$

$$\begin{aligned}P\{|\theta_n - \theta| > \varepsilon n^{-\delta}\} &\leq P\{|\Delta_n(\theta_n) - \Delta_n(\theta)| \geq K_0' \varepsilon n^{-\delta}\delta_n\} \\ &\leq P\{|\Delta_n(\theta_n) - \Delta_n(\theta)| n^\delta \geq K_0' \varepsilon c\} + P\{\delta_n < c\}.\end{aligned}$$

Since Lemma 3 applies to the first probability on the r.h.s. of this last inequality, it suffices to consider

$$\begin{aligned}P\{|\delta_n - \mu| > c\} &\leq P\{|\underline{\delta}_n[\theta n]/n - \theta\mu_F| > c/2\} \\ &\quad + P\{|\bar{\delta}_n(1 - [\theta n]/n) - (1-\theta)\mu_G| > c/2\}.\end{aligned}$$

The two probabilities in this upper bound are both handled in the same way; we only deal explicitly with the first one. It is dominated by

$$P\{\underline{\delta}_n|[\theta n]/n - \theta| > c/4\} + P\{\theta|\underline{\delta}_n - \mu_F| > c/4\}.$$

Of these two probabilities, the first is 0 for n sufficiently large (since $\underline{\delta}_n \le 1$ always holds). Since $\{\delta_{ni}^{\theta}: 1 \le i \le [\theta n]\}$ are iid and bounded, we can use (2.3) of Hoeffding (1963) to dominate the second probability by $2\exp\{-2[\theta n](c/4\theta)^2\}$. This bound can now be absorbed into the earlier bound from Lemma 3. □

Acknowledgment. I sincerely thank one referee for providing many insightful comments and recommendations. Specifically, the inclusion of a weighting factor in the criterion function, the use of Cramér–von Mises and Kolmogorov–Smirnov distances, the analysis of the Lindisfarne data and the consideration of invariance are all due to that referee.

REFERENCES

BHATTACHARYYA, G. K. and JOHNSON, R. A. (1968). Non-parametric tests for shift at an unknown time point. *Ann. Math. Statist.* **39** 1731–1743.

COBB, G. W. (1978). The problem of the Nile: Conditional solution to a change-point problem. *Biometrika* **65** 243–251.

DARKHOVSHK, B. S. (1976). A non-parametric method for the a posteriori detection of the "disorder" time of a sequence of independent random variables. *Theory Probab. Appl.* **21** 178–183.

DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642–669.

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

HINKLEY, D. V. (1972). Time-ordered classification. *Biometrika* **59** 509–523.

HINKLEY, D. V. and HINKLEY, E. A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika* **57** 477–488.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

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

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 Estadist. Investigación Oper.* **31** 83–98.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514