M. Baron. Nonparametric adaptive change-point estimation
and on-line detection. *Sequential Analysis*, 19 (1&2), 1-23, 2000.

Nonparametric adaptive change-point estimation and on-line detection

By Michael I. Baron
Programs in Mathematical Sciences, University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.
e-mail: mbaron@utdallas.edu

Summary

Under standard conditions of change-point problems with one or both distributions being unknown, we propose on-line and off-line nonparametric algorithms for detecting and estimating the change-point, based on histogram density estimators. Their asymptotic behavior is similar to that of the most efficient procedures in the case of known distributions. The stopping rule achieves an asymptotically linear mean delay and an exponential mean time between false alarms. The proposed procedures are applied to different data sets to detect the global climate changes.

Some key words: change-point, generalized likelihood ratio, Kullback-Leibler information, mean delay, mean time between false alarms, stopping time.

1. Introduction

It is often necessary to detect and to estimate a change-point, when distributions are unknown to a statistician. In many applications, one may know the distribution *before* the change, when the process is "in control", but it is usually impossible to know the distribution *after* the change, when the process goes "out of control". We propose retrospective and sequential (off-line and on-line) change-point estimation algorithms, when both distributions, or only the post-change distribution, are unknown.

Besides being nonparametric, the proposed algorithms possess the following three important properties. They are *universal* because they are designed to detect any change in distribution (and not just mean and variance shifts). Also, they may be applied to numerical, ordinal and categorical data, and the researcher does not have to

decide about the type of data. Finally, they are *adaptive* in that asymptotically they behave similarly to their parametric prototypes in the case of known distributions.

One observes a sample or a sequence $\{x_j\}$, where x_j has a distribution F for $j \le \nu$ and a distribution G otherwise. The change-point parameter ν is to be estimated from these data. When both distributions are known, many popular change-point estimation procedures are based on a log-likelihood ratio random walk

$$S_k = \sum_{j=1}^{k} \log \frac{f}{g}(x_j). \tag{1}$$

In the retrospective setting, Hinkley (1970) proposes a maximum likelihood estimator (MLE) $\hat{\nu} = \arg\max_k(S_k)$. Since $|\hat{\nu} - \nu| = O_p(1)$, as $n \to \infty$, and there is no consistent estimator, it is the best rate among all off-line procedures. We derive a nonparametric estimator with the same rate.

In the sequential setting, a sequence $x_1, x_2, \ldots$ is being collected. An on-line change detection algorithm is a stopping time T, which signals a change in distribution "as soon as possible" after it occurs. The stopping time $T(h)$, proposed by Page (1954) and obtained from Wald's sequential probability ratio test, is defined as $T(h) = \inf\{n : W_n \ge h\}$, where h is a chosen threshold, and W_k is the CUSUM process generated by S_k. Optimality of $T(h)$ in terms of the mean delay (MD) and the mean time between false alarms (MTBFA) is shown in Pollak (1985), Moustakides (1986) and Ritov (1990). Basseville and Nikiforov (1993), section 5.2, prove that

$$\text{MD} \le \frac{h + C_1}{\mathcal{K}(G, F)} \quad \text{and} \quad \text{MTBFA} \ge \frac{e^h - h - C_2}{\mathcal{K}(F, G)}, \tag{2}$$

where $\mathcal{K}(\cdot, \cdot)$ is the Kullback-Leibler information. We propose a nonparametric on-line change-point detection algorithm, whose mean delay is also at most linear in h, and whose mean time between false alarms is also at least exponential in h.

Both procedures, CUSUM and MLE, are well defined in the case of known distributions. If distributions belong to a known family, whose unknown (nuisance) parameter changes from θ_0 to θ_1, one estimates θ_0 and θ_1 from subsamples $(x_1, \ldots, x_k)$ and $(x_{k+1}, \ldots, x_n)$, respectively, for $k = 1, \ldots, n$. The estimates $\hat{\theta}_0(k)$ and $\hat{\theta}_1(k)$ are then substituted in (1) to result in the *generalized log-likelihood ratio* process (see Siegmund (1988) for Normal distributions, Hu and Rukhin (1995) for exponential families).

We extend this idea to the case of completely unknown distributions. That is, we estimate pre-change and post-change *densities* for every possible value of a change-point,

substitute in (1) and obtain nonparametric MLE-type and CUSUM-type procedures. Their performance is evaluated in retrospective setting (section 2), in sequential setting with the unknown post-change distribution (section 3), and with both distributions being unknown (section 4). Section 5 contains analysis of real and simulated examples.

2. An off-line procedure

Several nonparametric off-line change-point estimation procedures have been proposed in the literature. It is usual to compare pre- and post-change empirical distributions for every $k = 1, \dots, n$. Then the change-point is estimated by the point which tells them the farthest apart in some "metric" (or measure of diversity). In Carlstein (1988), a mean-dominant norm is used as such a "metric", and the rate $|\hat{\nu} - \nu| = O_p(n^{1/2+\epsilon})$ is achieved. Dümbgen (1991) uses a seminorm and achieves the rate $O_p(1)$, similarly to the MLE for the case of known distributions. Ferger's (1991) estimator has the rate $O_p(\log n)$. Our nonparametric MLE-type estimator is shown to maximize the (weighted) Kullback-Leibler information between estimated densities, which plays the role of a "metric". It has the unbeatable rate of $O_p(1)$.

We use the histogram density estimation procedures to estimate f and g (Devroye and Györfi, 1985). Smooth density estimators (e.g. kernel estimators) may be used along the same lines. We consider the histogram density estimators because they allow to detect changes in the distribution of categorical and ordinal data as well. Also, the resulting process $\hat{S}_k$ has a clear interpretation given by Propositions 1 and 2.

Let $\mathcal{X}$ be the united support of F and G with a measure μ, and let $\{A_m\}$ be its μ-measurable partition of rank r. If supports are not known, one can take an intentionally larger set $\mathcal{X}$. For $k = 1, \dots, n$, define

$$\begin{aligned} \hat{f}(x) &= \frac{1}{k\mu(A_m)} \sum_{m=1}^{r} \sum_{j=1}^{k} I\{x \in A_m, x_j \in A_m\}, \\ \hat{g}(x) &= \frac{1}{(n-k)\mu(A_m)} \sum_{m=1}^{r} \sum_{j=k+1}^{n} I\{x \in A_m, x_j \in A_m\}. \end{aligned} \tag{3}$$

The reference measure μ is chosen arbitrarily, because it does not enter the likelihood ratios. We substitute the estimates (3) in (1) and propose an *MLE-type change-point estimator* $\hat{\nu} = \arg\max_k \hat{S}_k$, where

$$\hat{S}_k = \sum_{j=1}^{k} \log \frac{\hat{f}(x_j)}{\hat{g}(x_j)} = \sum_{j=1}^{k} \log \frac{\sum_{i=1}^{k} \delta(\xi_i, \xi_j)/k}{\sum_{i=k+1}^{n} \delta(\xi_i, \xi_j)/(n-k)}, \tag{4}$$

$\xi_j^{(m)} = I\{x_j \in A_m\}$, for $m = 1, \ldots, r, j = 1, \ldots, n$, and $\delta(x, y) = I\{x = y\}$.

Using the histogram density estimators, one reduces the problem to estimating a change-point for the multinomial distribution of ξ_j. As shown below, $\hat{S}_k$ is proportional to the Kullback-Leibler information between estimated multinomial distributions, which suggests, in general, to maximize the Kullback-Leibler information between empirical pre- and post-change distributions.

Proposition 1 *One has* $\hat{S}_k = k \cdot \mathcal{K}(\hat{p}_{0:k}, \hat{p}_{k:n})$, *where* $\hat{p}_{u:v} = (\hat{p}_{u:v}(1), \ldots, \hat{p}_{u:v}(r))$ *is a vector of sample proportions* $\hat{p}_{u:v}^{(m)} = \left\{\xi_{u+1}^{(m)} + \ldots + \xi_v^{(m)}\right\}/(v-u)$ *computed from a subsample* $x_{u+1}, \ldots, x_v$ *for* $0 \le u < v \le n$.

Proof: This equality follows directly from the representation (4). □

Let $p = (p(1), \ldots, p(r))$ and $q = (q(1), \ldots, q(r))$ be the pre-change and the post-change parameters of the multinomial distribution of ξ_j, so that $\hat{p}_{1:k}$ estimates p and $\hat{p}_{k+1:n}$ estimates q. The next proposition ties $\hat{S}_k$ with its parametric analogue Λ_k.

Proposition 2 *If* $\hat{S}_k$ *is the nonparametric estimator of the log-likelihood process* S_k *for estimating the change-point between distributions* F *and* G, *and* Λ_k *is the parametric generalized log-likelihood ratio process for estimating the change-point between* p *and* q, *then* $\hat{S}_k \equiv \Lambda_k$ *with probability one.*

Proof: From (4),

$$\frac{\hat{f}(x_j)}{\hat{g}(x_j)} = \frac{\sum_m \hat{p}_{0:k}^{(m)} \xi_j^{(m)}}{\sum_m \hat{p}_{k:n}^{(m)} \xi_j^{(m)}} = \prod_m \left(\frac{\hat{p}_{0:k}^{(m)}}{\hat{p}_{k:n}^{(m)}}\right)^{\xi_j^{(m)}},$$

which is the likelihood ratio for the multinomial distribution. □

Hence, $\hat{\nu}$ coincides with its parametric prototype for the multinomial case. Next, the distribution of ξ_j can be embedded in an $(r-1)$-parameter minimal standard exponential family (Brown (1986), chapter 1). Then, $|\hat{\nu} - \nu| = o_p(n^\alpha)$ if $\min\{\nu, n-\nu\} \ge ln^\omega$ for some $l > 0$ and $\omega > \alpha$ (Proposition 3.1 of Baron and Rukhin (1997) with $c = 0$). A stronger statement holds if ν/n is bounded away from 0 and 1.

Theorem 3 *Suppose that*

$$\begin{aligned}
&(i) \quad \liminf_{n\to\infty} \frac{1}{n} \min\{\nu, n-\nu\} > 0, \\
&(ii) \quad (p(1), \ldots, p(r)) \neq (q(1), \ldots, q(r)), \\
&(iii) \quad \rho = \min_{m=1\ldots r} \min\{p^{(m)}, q^{(m)}\} > 0.
\end{aligned}$$

Then $|\hat{\nu} - \nu| = O_p(1)$ *as* $n \to \infty$.

Note that this rate of $|\hat{\nu} - \nu|$ can not be improved, because it is the optimal rate achieved in the case of known distributions (see Hinkley (1970)).

Condition (i) is weaker than the standard assumption $\nu/n \to \tau \in (0;1)$ used in Carlstein (1988) and Dümbgen (1991).

Failure to satisfy (ii) yields equality of the pre-change and post-change probabilities of A_m, i.e. there is no change in distribution in terms of this partition. Any information about the experiment will assist in choosing the partition. For example, a partition of rank $r = 2$, consisting of $A_1 = (-\infty, x_0]$ and $A_2 = (x_0, +\infty)$, satisfies (ii) in the case of any mean shift of a continuous distribution for any $x_0 \in \mathrm{supp}(F) \bigcup \mathrm{supp}(G)$.

Condition (iii) implies that the partition rank should not be too high. Usually, one needs no more than 5 partition sets to tell two distributions apart. If supports of F and G are known, one can always satisfy condition (iii). If the distributions are supported by the whole real line, this condition is automatically satisfied.

Problems in numerical applications are caused by the fact that histogram density estimators may be equal 0, which yields $\hat{S}_k = \pm\infty$. If $\hat{p}^{(m)}_{0:k} = 0$ for some m, we let $\hat{p}^{(m)}_{0:k} \log\left\{\hat{p}^{(m)}_{0:k}/\hat{p}^{(m)}_{k:n}\right\} = 0$. To avoid $\hat{p}^{(m)}_{k:n} = 0$, one can use the smoothed version. Let s be the number of zeros among $\hat{p}_{k:n}(1), \ldots, \hat{p}_{k:n}(r)$. Then, consider

$$\tilde{p}^{(m)}_{k:n} = \begin{cases} \hat{p}^{(m)}_{k:n} & \text{if } s = 0 \\ \varepsilon/s(n-k) & \text{if } \hat{p}^{(m)}_{k:n} = 0 \\ (1 - \varepsilon/(n-k))\,\hat{p}^{(m)}_{k:n} & \text{if } s > 0;\ \hat{p}_{k:n} \neq 0 \end{cases} \tag{5}$$

for some $0 < \varepsilon < 1$. This definition still keeps $\sum_m \tilde{p}^{(m)}_{k:n} = 1$. The probability that $\tilde{p}_{k:n} \neq \hat{p}_{k:n}$ for some k within n^α units from n converges to 0, hence, the statement of Theorem 3 remains valid with $\hat{p}_{k:n}$ replaced by $\tilde{p}_{k:n}$.

3. An on-line algorithm for the known pre-change distribution

Sections 3 and 4 deal with data collected sequentially in time. An optimal stopping time T must achieve a reasonable tradeoff between minimizing the mean delay

$$\mathrm{MD}(T) = \mathrm{E}_\nu (T - \nu)^+ \tag{6}$$

and maximizing the mean time between false alarms

$$\mathrm{MTBFA}(T) = \mathrm{E}_{\nu=\infty}(T) = \mathrm{E}\{T | \text{no change}\}.$$

We propose a nonparametric analogue of the CUSUM algorithm, whose asymptotic behavior is similar to that in the case of known distributions. In this section, we assume that the pre-change distribution F is known, whereas the post-change distribution G is completely unknown. Then, its density in (1) is again replaced by a histogram density estimator (3), and the CUSUM process is estimated by

$$\hat{W}_n = \max_{1 \le k < n} \hat{S}_{k:n},$$

where

$$\hat{S}_{k:n} = (n-k) \sum_{m=1}^{r} \hat{p}_{k:n}^{(m)} \log \frac{\hat{p}_{k:n}^{(m)}}{p^{(m)}}. \tag{7}$$

is the "tail" of the generalized log-likelihood ratio process. Next, we choose a positive threshold h and propose a stopping time

$$\hat{T}(h) = \inf \left\{ n : \ \hat{W}_n \ge h \right\}. \tag{8}$$

Theorem 4 *Under assumptions (ii) and (iii) of Theorem 3, one has for the stopping time* $\hat{T}(h)$,

$$MTBFA \ge \frac{2}{3(Be^{A/B})^{1/2}} e^{h/B} - \frac{2}{3} \tag{9}$$

and

$$MD \le \frac{h}{\mathcal{K}(p,q)} + \left(\frac{8hr}{(1-a)\mathcal{K}(q,p)} \right)^{1/2} + \frac{2r}{1-a}, \tag{10}$$

where

$$A = \sum_m \frac{\log p^{(m)}}{p^{(m)}}, \ B = \sum_m \frac{1}{p^{(m)}}, \ a = \exp\left\{ -\frac{2\mathcal{K}^2(p,q)}{\mathcal{D}^2(p,q)} \right\}, \ \mathcal{D}(p,q) = \sum_m \left| \log \frac{q^{(m)}}{p^{(m)}} \right|. \tag{11}$$

The order of MTBFA in (9) is different from that in (2). Indeed, under the no-change assumption, one has no data from the distribution G. If G is unknown, then the lower bound in (9) must be independent of it, hence, it has to differ from (2).

4. An on-line algorithm for the unknown pre-change distribution

As we show below, it gets fundamentally more difficult to estimate the change-point if the pre-change distribution is not available. Several problems arise in this case.

If both distributions are unknown, one has to observe the process for a sufficiently long period *before* ν. Indeed, for any finite interval $[s;t]$, if f is estimated

from $(x_s, \dots, x_t)$, then with a positive probability $\hat{f} \approx g$, and no change can be detected. Hence, one should not compare $\hat{g}$ with $\hat{f}$ for small k. We address this issue by redefining the CUSUM-type process and letting

$$\hat{W}_n = \max_{\gamma n \le k < n} \hat{S}_{k:n} \text{ for some } \gamma \in (0; 1), \tag{12}$$

where γn may be replaced by a nonlinear function $\gamma(n)$. This reflects the fact that no algorithm can detect *frequent* distributional changes, when both distributions are unknown. Here, changes can be reported at most once per every γn observations.

Then, at least γn observations are used to estimate $f(\cdot)$. On the other hand, if a change at the time ν is not detected before the time ν/γ, then for all $k > \nu/\gamma$ the density f is estimated from a sample, at least $100(1 - \nu/\gamma n)$ percent of which comes from the distribution G. This portion increases and tends to 100% as $n \to \infty$, which makes it impossible to detect the change-point if it occurred far in the past.

Hence, for $n >> \nu$, $\hat{W}_n$ behaves like in the no-change situation, when all data follow the distribution G. Therefore, if a change-point is not detected before ν/n, $\hat{T}(h)$ will increase exponentially fast in h, like the mean time between false alarms.

Thus, an important characteristic of a stopping time is $\text{pr}\{T > \nu/\gamma\}$, which is the probability to fail to detect the change-point. For the proposed algorithm, this probability is exponentially small, when n is large. Also, it is reasonable to study

$$MD' = \text{E}\left\{(\hat{T} - \nu)^+ I\{T < \nu/\gamma\}\right\},$$

or $\text{E}\left\{(T - \nu)^+ | T < \nu/\gamma\right\}$, instead of an unconditional mean delay (6).

To implement the proposed procedures, one defines $\tilde{p}_{0:k}$ similarly to (5), guaranteeing $\tilde{p}^{(m)}_{0:k} > 0$ and $\sum_m \tilde{p}^{(m)}_{0:k} = 1$. Then, $p^{(m)}$ is replaced by $\tilde{p}^{(m)}_{0:k}$ in (7), a truncated CUSUM-type process $\hat{W}_n$ is introduced by (12), and a stopping time $\hat{T}(h)$ is defined by (8). The next theorem evaluates the performance of $\hat{T}(h)$ in this case.

Theorem 5 *Under assumptions (ii) and (iii) of Theorem 3,*

$$\begin{aligned} MTBFA \;\; \ge \;\; & \frac{2}{3(2r - 2r\gamma)^{1/2}} \exp\left\{\frac{h}{2B\{1 + (1/\gamma - 1)^{1/2}\}^2}\right\} \times \\ & \times \left(1 + O(\exp\{-\frac{bh}{\log h}\})\right), \quad as \;\; h \to \infty, \end{aligned} \tag{13}$$

$$MD' \le \left\{\frac{h}{\mathcal{K}(p,q)} + \left(\frac{8hr}{(1-a)\mathcal{K}(p,q)}\right)^{1/2} + \frac{2r}{1-a}\right\}(1 + O(1/\nu)), \tag{14}$$

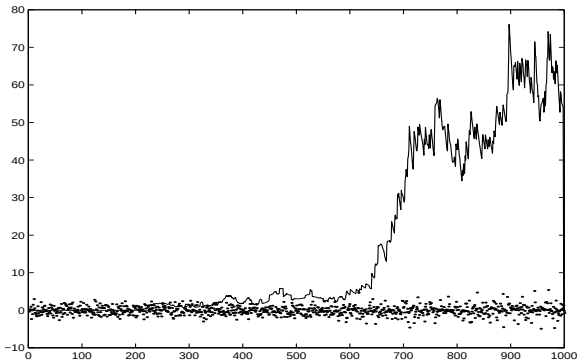

Figure 1: *Retrospective (left) and sequential (right) change-point detection.*

as $\nu \to \infty$, *and*

$$\mathrm{pr}\left\{\hat{T}(h) > \nu/\gamma\right\} \le 2ra^{(1/\gamma-1)\nu}\left\{1 + o(\nu^{-1/2})\right\} \to 0, \quad as \ \ \nu \to \infty, \tag{15}$$

where A, B *and* a *are defined by (11), and* $b = \gamma \min_m p^{(m)}/(2-2\gamma)$.

5. Applications

We assume that pre-change and post-change distributions are completely unknown. If a researcher possesses any information about F or G, the proposed algorithms are modified by using the best available estimates instead of (3). For example, if F belongs to a known family, then $\hat{f}(x_j)$ is replaced by $f(x_j|\hat{\theta})$. Clearly, these algorithms will perform at least as well as stated in Theorems 3, 4 and 5.

We first look at the performance of the proposed procedures using simulated data. On Figure 1 (left), the distribution of the data undergoes a small change from Normal(0,1) to Normal ($\frac{3}{10}$,1) at the time $\nu = 600$. The generalized log-likelihood ratio process detects the change at $\hat{\nu} = 592$. Four zones with different theoretical probabilities $p^{(m)}$ are used, $A_1 = (-\infty, -2]$, $A_2 = (-2, 0]$, $A_3 = (0, 2]$ and $A_4 = (2, +\infty)$.

Sequential change detection procedure $\hat{T}(h)$ is applied to the simulated data set, where the standard Normal distribution changes to standard Laplace at $\nu = 600$ (Figure 1, right). Note that both distributions are symmetric, with equal expected values and variances. Of course, the result depends on a threshold h. For example, $\hat{T}(40) = 708$, $\hat{T}(20) = 678$, $\hat{T}(10) = 647$, $\hat{T}(6) = 622$, and $\hat{T}(5) = 460$, i.e. with $h \le 5$ the algorithm reports a false alarm.

Next, we apply the proposed algorithms to famous data sets. The problem of the Nile (Cobb (1978)) concerns the change in the annual volume of the Nile river near the

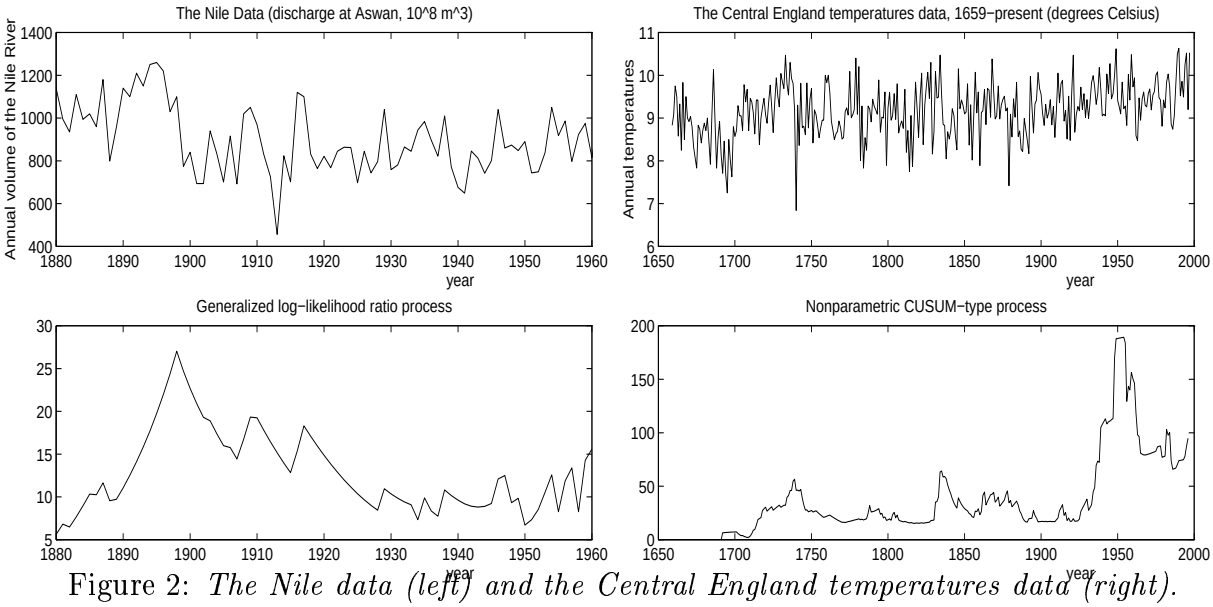

Figure 2: *The Nile data (left) and the Central England temperatures data (right).*

city of Aswan, Egypt. We use the off-line nonparametric procedure $\hat{\nu}$ to estimate the change-point. The generalized log-likelihood ratio process $\hat{S}_k$ is depicted on Figure 2 (left) along with the data. Scientists believe that the change occurred in 1898. This is supported by various estimators in Cobb (1978), Carlstein(1988) and Dümbgen(1991). Our nonparametric algorithm $\hat{\nu}$ detects a change-point in 1898 too. We used $r = 3$ zones, $A_1 = (-\infty, 850)$, $A_2 = [850; 950)$ and $A_3 = [950, +\infty)$.

The other data set (Figure 2, right) represents the average annual temperatures every year from 1659 till 1997 in the area of the United Kingdom enclosed by Preston, London and Bristol (see Manley, 1974, updated by Parker, 1992, and the Climate Data Monitoring section of the Hadley Centre). For lower values of the threshold h (between 30 and 50), the climate changes are detected around 1740 and 1840. These years correspond to the maximum development and the end of the Little Ice Age. However, any threshold between 30 and 190 detects a significant change point before the year 1950. This climate change is likely to be related to a so-called greenhouse effect caused by the increasing concentration of carbon dioxide and other trace gases in the atmosphere.

Earlier climate changes can be detected from the famous data set of atmospheric temperatures over 160,000 years, obtained from the Vostok, Antarctica, ice core (Jouzel et al, 1987). During this period, the climate changed more than once. The off-line change-point estimator $\hat{\nu}$ points to the time 143,000 years ago, which is perhaps the

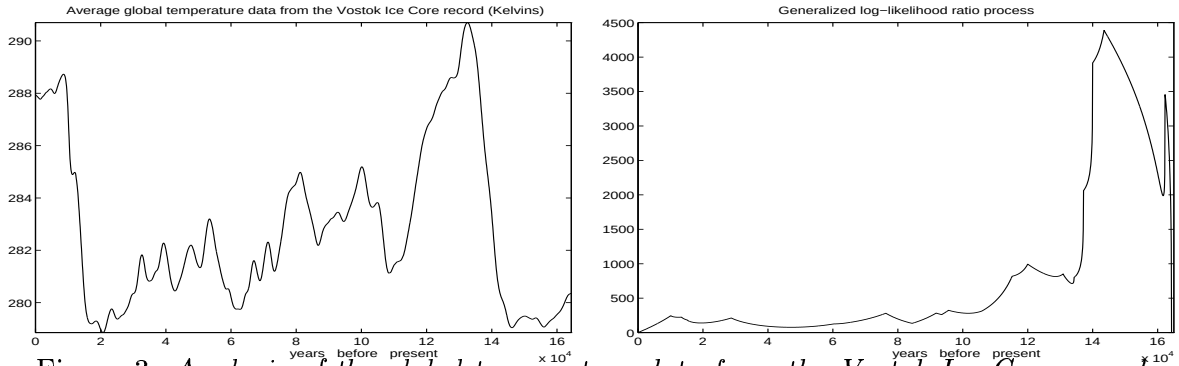

Figure 3: *Analysis of the global temperature data from the Vostok Ice Core record.*

most significant climate change (Figure 3). It is the time of the last interglacial phase.

Appendix

Lemma 1 *Let ζ_n be a sequence of random variables. Then $\zeta_n = O_p(1)$ if and only if $\zeta_n = O_p(g_n)$ for any unboundedly increasing sequence g_n.*

Proof: If $\zeta_n \neq O_p(1)$, then there exist $\varepsilon > 0$ and a sequence $\{\lambda_j\}, \lambda_j \to \infty$, so that $\limsup_{n\to\infty} \mathrm{pr}\{\zeta_n > \lambda_j\} > \varepsilon$ for all $j \geq 1$. Hence, $\mathrm{pr}\left\{\zeta_{n(k)} > \lambda_j\right\} > \varepsilon$ for some subsequence $\zeta_{n(k)}$. In particular, $\mathrm{pr}\left\{\zeta_{n(k)} > \lambda_{n(k)}\right\} > \varepsilon$. Hence, $\zeta_n \neq o_p(\lambda_n)$, which implies that $\zeta_n \neq O_p(g_n)$ for $g_n = \lambda_n^{1/2}$. The other implication is trivial. □

Outline of the Proof of Theorem 3

Choose $0 < \alpha < 1$ and a sequence $g_n \to \infty$, $g_n < n^\alpha$. Using Proposition 1 and convex properties of $\mathcal{K}(\cdot,\cdot)$, one has for any $k \in [\nu - n^\alpha; \nu - g_n]$,

$$\hat{S}_\nu - \hat{S}_k = (\nu - k)\mathcal{K}(p,q) + k(\mathcal{K}(p,q) - \mathcal{K}(p,\tilde{q})) + R \geq (\nu - k)\mathcal{K}(p,q) + R, \qquad (16)$$

where $\tilde{q} = p(\nu - k)/(n - k) + q(n - \nu)/(n - k)$ and R is the remainder,

$$R = \nu\left[\mathcal{K}(\hat{p}_{0:\nu}, \hat{p}_{\nu:n}) - \mathcal{K}(p,q)\right] - k\left[\mathcal{K}(\hat{p}_{0:k}, \hat{p}_{k:n}) - \mathcal{K}(p,\tilde{q})\right].$$

The Taylor expansion of $x \log x$ and $\log x$ for $x = \hat{p}_{0:k}, \hat{p}_{0:\nu}, \hat{p}_{k:n}$ and $\hat{p}_{\nu:n}$ about their respective expected values $p, p, \tilde{q}$ and q shows that $R = O_p\left\{(\nu - k)^{1/2}\right\}$, as $n \to \infty$. Hence, from (16) and the Hoeffding's lemma (Hoeffding, 1963),

$$\mathrm{pr}\left\{\hat{S}_\nu \leq \hat{S}_k\right\} \leq \mathrm{pr}\{|R| \geq (\nu - k)\mathcal{K}(p,q)\} \leq \alpha \exp\{-\beta(\nu - k)\}$$

for some $\alpha, \beta > 0$, and a similar inequality holds for $k > \nu + g_n$. Hence,

$$\mathrm{pr}\{|\hat{\nu} - \nu| > g_n\} = \mathrm{pr}\left\{\bigcup_{|k-\nu|>g_n} \{\hat{S}_{\hat{\nu}} \le \hat{S}_k\}\right\} \le \sum_{k:|k-\nu|>g_n} a \exp\{-b(\nu - k)\} \to 0.$$

Since g_n was chosen arbitrarily, the theorem follows from Lemma 1.

Outline of the Proof of Theorem 4

1). Let $\{u_m,\ m = 1, \ldots, r\}$ be such that $\sum_m u_m^2/p^{(m)} \le h/(n-k)$ and $u_m > 0$. If all observations come from the distribution F, then

$$\mathrm{pr}\left\{\hat{S}_{k:n} \ge h\right\} \le \mathrm{pr}\left\{\bigcup_m \left(|\hat{p}_{k:n}^{(m)} - p_{k:n}^{(m)}| \ge u_m\right)\right\} \le 2\sum_m \exp\left\{-2(n-k)u_m^2\right\} \qquad (17)$$

by the Hoeffding's lemma. The right-hand side of (17) can be replaced by an infimum over $\{u_m\}$, which equals $B\exp\{(A-2h)/B\}$ for all

$$h > \max_m \left\{A - B\log p^{(m)}\right\}. \qquad (18)$$

Hence, for any integer M,

$$\begin{aligned}\mathrm{E}\left\{\hat{T}(h)\right\} &= \sum_{N=1}^{\infty} \mathrm{pr}\left\{\hat{T} > N\right\} \ge \sum_{N=1}^{M}\left(1 - \sum_{n=2}^{N}\sum_{k=1}^{n-1} \mathrm{pr}\left\{\hat{S}_{k:n} \ge h\right\}\right) \\ &\ge \sum_{N=1}^{M}\left\{1 - N(N-1)Be^{A/B}e^{-2h/B}\right\}\end{aligned}$$

Choosing $M = \left[e^{h/B}(Be^{A/B})^{-1/2}\right]$, one obtains (9) for all h satisfying (18). For all other h,

$$\frac{e^{h/B}}{(Be^{A/B})^{1/2}} \le \frac{\exp\{(A/B - \log p^{(m)})/2\}}{(Be^{A/B})^{1/2}} = \left(\frac{1/p^{(m)}}{\sum_l 1/p^{(l)}}\right)^{1/2} \le 1.$$

which nullifies the right-hand side of (9). Thus, (9) holds for any $h > 0$.

2). Without loss of generality, assume $\nu = 0$, i.e. all observations are coming from the distribution G. Then, similarly to the proof of Theorem 3, one writes $\hat{S}_N = N\{\mathcal{K}(q,p) + R_N\}$ and shows that $\mathrm{pr}\left\{\hat{S}_N < h\right\} \le 2ra^{V(h,N)}$ for any $N > h/\mathcal{K}(q,p)$, where $V(h,N) = N - 2h/\mathcal{K}(q,p) + h^2/N\mathcal{K}^2(q,p)$, and a is defined by (11).

Then one estimates $\sum a^{V(h,N)}$ from above and obtains,

$$\mathrm{E}_{\nu=0}\left\{\hat{T}(h)\right\} = \sum_{N=1}^{\infty}\left\{\hat{T}(h) > N\right\} \le \sum_{N=1}^{\infty} \mathrm{pr}\left\{\hat{S}_N < h\right\} \le \left(1 + \frac{1}{n}\right)\frac{h}{\mathcal{K}(q,p)} + \frac{2rn}{1-a} \qquad (19)$$

for any $n \ge 1$. Minimizing the right-hand side of (19) in n leads to (10).

Outline of the Proof of Theorem 5

1). Let $c = B^{-1}\{1 + (1/\gamma - 1)\}^{-2}$, and consider the event

$$\mathcal{E}_k = \left\{2\tilde{p}_{0:k}^{(m)} > p^{(m)},\ 2k(\tilde{p}_{0:k}^{(m)} - p^{(m)})^2 < ch,\ 2(n-k)(\hat{p}_{k:n}^{(m)} - p^{(m)})^2 < ch,\ m = 1 \dots r\right\}.$$

If $\mathcal{E}_k$ occurs, then $\hat{S}_{k:n} < h$. Therefore, by the Hoeffding's lemma,

$$\mathrm{pr}\left\{\hat{S}_{k:n} \geq h\right\} \leq 1 - \mathrm{pr}\left\{\mathcal{E}_k\right\} \leq 4r \exp\left\{-ch\right\} + 2r \exp\left\{-c'k\right\},$$

where $c' = \min_m (p^{(m)})^2/2$. Also, $\mathrm{pr}\left\{\hat{S}_{k:n} \geq h\right\} = 0$ if $n \leq h/(1-\gamma)\log h = \phi(h)$ and $h > \exp\{(r-1)\gamma/\varepsilon(1-\gamma)\}$. Hence,

$$\begin{aligned} MTBFA &\geq \sum_{N=1}^{M} \left\{1 - \sum_{n=\phi(h)}^{N} \sum_{k=\gamma n}^{n-1} \mathrm{pr}\left\{\hat{S}_{k:n} \geq h\right\}\right\} \\ &\geq \sum_{N=1}^{M} \left\{1 - 2r\left(2(1-\gamma)N(N-1)e^{-ch} + \frac{e^{-c'\gamma\phi(h)}}{(1-e^{-c'})(1-e^{-c'\gamma})}\right)\right\} \end{aligned}$$

for all M. The choice of $M = \left[\exp\left\{ch/2\right\}(2r - 2r\gamma)^{-1/2}\right]$ leads to (13).

2). The mean delay. According to Theorem 4, under fixed $\left\{\tilde{p}_{0:k}^{(m)}, m = 1 \dots r\right\}$,

$$\mathrm{E}\left\{MD' | \tilde{p}_{0:k}(1), \dots, \tilde{p}_{0:k}(r)\right\} \leq \frac{h}{\mathcal{K}(q, \tilde{p}_{0:k})} + \left(\frac{8hr}{(1-\tilde{a})\mathcal{K}(q, \tilde{p}_{0:k})}\right)^{1/2} + \frac{2r}{1-\tilde{a}}, \qquad (20)$$

where $\tilde{a} = \exp\left\{-2\left\{\mathcal{K}(q, \tilde{p}_{0:k}) / \sum_m |\log(q/\tilde{p}_{0:k})|\right\}^2\right\}$. One shows that

$$\begin{gathered} \mathrm{E}\left(\tfrac{1}{\mathcal{K}(q,\tilde{p}_{0:k})}\right) \leq \tfrac{1}{\mathcal{K}(q,p)} + O(1/\nu), \qquad \mathrm{E}\left(\tfrac{1}{1-\tilde{a}}\right) = \tfrac{1}{1-a}(1 + O(1/\nu)), \\ \mathrm{E}\left\{(1-\tilde{a})\mathcal{K}(q,\tilde{p})\right\}^{-1/2} \leq \left\{(1-a)\mathcal{K}(q,p)\right\}^{-1/2}(1 + O(1/\nu)), \end{gathered}$$

as $\nu \to \infty$, and obtains (14) by taking expected values in (20).

3). Probability of failure to detect the change-point by the time ν/γ. Similarly to the proof of (14), condition on $x_j,\ j \leq \nu$, and obtain,

$$\mathrm{pr}\left\{\hat{T}(h) > \nu/\gamma\right\} \leq \mathrm{E}\left\{\mathrm{pr}\left(\hat{S}_{\nu:[\nu/\gamma]} < h | x_1, \dots, x_\nu\right)\right\} \leq 2r\mathrm{E}e^{-2C\nu},$$

where $C = (1/\gamma - 1)\left\{\mathcal{K}(q,\tilde{p}) - h/(\nu/\gamma - \nu)\right\}^2 / \mathcal{D}^2(\tilde{p}, q)$, and $\mathcal{D}$ is defined in (11). Then (15) follows because both $\mathcal{K}(q,\tilde{p}) - \mathcal{K}(q,p)$ and $\mathcal{D}(\tilde{p},q) - \mathcal{D}(p,q)$ are $O_p(\nu^{-1/2})$.

References

Baron, M. & Rukhin, A. L. (1997). Asymptotic behavior of confidence regions in the change-point problem. *J. Statist. Planning Inference* **58**, 263–282.

Basseville, M. & Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application.* PTR Prentice-Hall, Inc.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Lecture Notes–Monograph Series, Volume 9, Hayward, CA.

Carlstein, E. (1988). Nonparametric estimation of a change-point. *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.

Devroye, L. Györfi, L. (1985). *Nonparametric density estimation. The L_1 view.* Wiley, New York.

Duembgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann. Statist.* **19**, 1471–1495.

Ferger, D. (1991). *Nonparametric change-point detection based on U-statistics.* PhD thesis, University of Giessen.

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

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

Hu, I. & Rukhin, A. L. (1995). A lower bound for error probability in change-point estimation. *Statistica Sinica* **5**, 319–331.

Jouzel, J., C. Lorius, J.R. Petit, C. Genthon, N.I. Barkov, M. Kotlyakov & M. Petrov (1987). Vostok ice core: a continuous isotope temperature record over the last climatic cycle (160,000 years). *Nature* **329**, 403–408.

Manley, G. (1974). Central England Temperatures: monthly means 1659 to 1973. *Q.J.R. Meteorol. Soc.* **100**, 242–261.

Moustakides, G. V. (1986). Optimal stopping times for detecting a change in distribution. *Ann. Statist.* **14**, 1379–1388.

Page, E. S. (1954). Continuous inspection schemes. *Biomterika* **41**, 100–115.

Parker, D. E., Legg, T. P., and Folland, C. K. (1992). A new daily Central England Temperature Series, 1772-1991. *Int. J. Clim.* **12**, 317–342.

Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13**, 206–227.

Ritov, Y. (1990). Decision theoretic optimality of cusum procedure. *Ann. Statist.* **18**, 1464–1469.

Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review* **56**, 31–48.