Consistent estimation of early and frequent change points

*

Michael I. Baron Nira Granott

March 14, 2005

Abstract

We address two types of processes with change points that often arise in practical situations. These are processes with *early change points* and processes with *frequent change points*. Early change points may occur after very few observations and may be followed by additional change points or more complicated patterns. Frequent change points separate different homogeneous phases of the observed process with the possibility of very short phases.

Uncertainty of the considered processes during their later phases forces the use of sequential tools, in order to minimize samples from later phases. Change-point detection and post-estimation schemes for these situations are developed. They possess a number of desired properties, not satisfied by procedures proposed in the earlier literature. One of them is distribution-consistency. Unlike the traditional concept of consistency, it implies convergence of small-sample change-point estimators to the corresponding parameters as the magnitude of changes tends to infinity.

1 Introduction and examples

In classical change-point problems, the distribution of observed data changes at an unknown moment, which is the parameter of interest. Sample $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ consists of two subsamples from distributions f and g, respectively, separated by an unknown change point ν. A vast amount of literature covers sequential and retrospective methods of change-point estimation. Classical references are eg. [7], [8], [9], [10], [13], [14], [22], [23], [25]; also see [5], [6], [15], and [28] for a survey of on-line and off-line algorithms.

Most of the proposed estimation schemes assume exactly one change point in the observed process and sufficiently large pre-change and post-change sample sizes. However, both assumptions are violated in a number of applications including developmental, cognitive, and educational psychology, energy pricing, meteorology, and quality control.

In problem solving processes, the first insight about a solution can occur after only a few solution attempts. Consequently, the first change point in the

distribution of solution times represents the moment of the first discovery and the end of a trial-and-error phase. It may occur after very few solutions, leading to a problem of change-point detection from small data sets ([3], see Fig. ??).

Typical processes of learning and development consist of several phases, separated by change points. Some phases are rather short. Fast learning, for example, is often associated with frequent changes of strategies ([2]). Fig. ?? presewnts results of an experiment where participants had to learn the functions of a given robot. Different strategies included observing the robot, turning it on and observing its performance, dismantling the robot and exploring its parts. Participants of the study used different strategies subsequently. The frequency of the alternation of strategies changed repeatedly, indicating different phases in the learning process. The times of a continuous use of the same strategy are depicted on Fig. ??. Periods of short single-strategy use times indicate phases of frequent alternation of strategies.

We define a process with an *early change point* to be a sequence of random variables $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots)$, where

$$\left\{ \begin{array}{lcll} \mathbf{X}_1 = (X_1, \ldots, X_\nu) & \sim & f, & \nu \geq 1 \\ \mathbf{X}_2 = (X_{\nu+1}, \ldots, X_\kappa) & \sim & g, & \kappa \geq \nu + 1 \\ \mathbf{X}_3, \mathbf{X}_4, \ldots, \ldots & \sim & ? & \end{array} \right. \tag{1}$$

After the first change point ν, observations come from distribution g until some unknown moment κ. After κ, no tractable model is assumed, as the later part of the process may contain further change points or even more complicated patterns and trends. The objective is to estimate the first change point ν.

A process with *multiple change points* will then be described as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_\lambda)$, where

$$\left\{ \begin{array}{lcl} \mathbf{X}_1 = (X_{\nu_0}, \ldots, X_{\nu_1}) & \sim & f_1 \\ \mathbf{X}_2 = (X_{\nu_1+1}, \ldots, X_{\nu_2}) & \sim & f_2 \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots & & \\ \mathbf{X}_\lambda = (X_{\nu_{\lambda-1}+1}, \ldots, X_{\nu_\lambda}) & \sim & f_\lambda \end{array} \right. \tag{2}$$

where $f_1, \ldots, f_\lambda$ are either known or unknown densities, ν_i, $i = 1, \ldots, \lambda - 1$, are change points (with a convention that $\nu_\lambda = N$ is the total sample size and $\nu_0 = 1$), and λ is the unknown number of homogeneous subsamples. If $|\nu_{k+1} - \nu_k|$ is small for some k, we will call this model a process with *frequent change points*.

Notice that densities f_i and f_j may coincide if $|i - j| \geq 2$. In the extreme case, all the odd subsamples come from the same distribution f_1, and all the even subsamples come from the same distribution f_2, so that the process oscillates between two densities f_1 and f_2. Such a model is suitable for detrended electricity prices, where relatively long "regular" or "control" periods alternate with short-term "spikes" ([4], [11], see Fig. ??).

The objective here is to estimate $(\lambda; \nu_1, \ldots, \nu_{\lambda-1})$, a parameter of an unknown dimension! The possibility of a homogeneous sample with no change points ($\lambda = 1$) is not excluded.

In both (1) and (2), the search for the first change point is complicated by uncertain behavior of the process after the end of the second phase. Thus, although the problems of estimating early or frequent change points are *retrospective* (not sequential) in general, it is not recommended to use retrospective statistical procedures, because they utilize the entire data set including the uncertain phases of the process. Conversly, it is desirable to use *sequential* tools that will utilize the minimum number of observations and stop as soon as the change point is detected.

Essentially, for the process with an early change point, one can resample one observation at a time until the first change is detected. In a situation with multiple change points, one will estimate the detected change point, discard all the pre-change observations, and search for the next change point sequentially. Therefore, both types of estimation problems, early change points and frequent change points, can be treated by similar techniques based on sequential detection and post-estimation of change points.

Many competing sequential schemes can be proposed. We select an optimal procedure according to the introduced principle of *distribution consistency*. It implies convergence of each change-point estimator to the corresponding parameter when sample sizes are fixed, but the magnitude of a change becomes more and more significant. In problems with early and frequent change points yielding short phases, we find this property at least as important as consistency in the classical sense.

The algorithm and its optimal properties are described in Section 2. In Section 3, we use the proposed scheme to detect possible global climate changes, spikes of electricity prices, and different phases in learning and problem solving processes.

2 Sequential estimation of multiple change points: a distribution consistent scheme

Several multiple change-point estimation schemes have been described in the literature. A maximum likelihood based procedure is proposed in [12]. However, especially in the case of unknown distributions, the naive maximum likelihood scheme is likely to return a change at every point, unless restrictions are enforced on the number of change points λ or the minimum distance Δ between them ([16], [17]). Still, this *restricted* maximum likelihood scheme can be shown to detect change points in any sample of $N > 2\Delta$ Bernoulli(p) variables if and only if at least two observations are different. It follows that the probability of a false alarm is as high as $1 - p^N - (1-p)^N$ in this case.

A conceptually different *binary segmentation* scheme ([27]) is an iterative procedure that divides the observed sample into two most distant subsamples, then divides each subsample, etc., until all the obtained subsamples are homo-

geneous. The disadvantage of this scheme is that no more than one change point is assumed at each step. For example, in the case of two intermittent distributions, as on Fig. **??**, it is unlikely to find a point separating two significantly different subsamples.

These problems can be resolved by a sequential estimation scheme that (1) considers increasing subsamples instead of the entire sample that may contain complicated patterns; (2) detects one change point at a time and does not assume its uniqueness in the observed data; (3) has an option of detecting 0 change points; (4) is sufficiently sensitive in order to detect a change point occurring after a short phase.

For the process with an early change point, the scheme consists of several steps outlined below. For the process with frequent change points, these steps are repeated until the entire data set is resampled.

2.1 Step 1: sequential detection

Observations are sampled sequentially until a stopping rule detects a change point. For the examples in Section 1, we used a stopping rule

$$T(h) = \inf \{n : W_n \geq h\},$$

based on a *cusum process*

$$W_n = \max_k \sum_{j=k+1}^{n} \log \frac{g(X_j)}{f(X_j)},$$

where f and g are the pre- and post-change densities or probability mass functions. Optimality of $T(h)$ is shown in [18] and [24]. In the case of unknown densities, one uses their best estimates for each "potential" value k of a change point, computes the *generalized likelihood ratio* based cumulative sums

$$\tilde{W}_n = \max_k \sum_{j=k+1}^{n} \log \frac{\hat{g}(X_j)}{\hat{f}(X_j)}, \tag{3}$$

and defines the stopping rule $\tilde{T}(h)$ similarly to $T(h)$. This stopping rule achieves asymptotically equivalent mean delay and mean time between false alarms ([1]).

Facing a possibility of early or frequent change points, one should increase sensitivity of the algorithm by choosing a low threshold h or a high probability of type I error α. The price to be paid is the increasing rate of false alarms, however, false change points will (hopefully) be eliminated at Step 3.

If only a sample of size N is available, all abovementioned stopping rules are curtailed so that $P\{T \leq N\} = 1$. In the case when $T(h) = N$ and $\tilde{W}_n < h$, the scheme results in zero detected change points. In all the other cases, a change point is detected and its location needs to be estimated.

2.2 Step 2: post-estimation

Notice that the stopping rule T itself is a poor estimator of the change point ν. Indeed, if $T \leq \nu$, it is a false alarm. If $T > \nu$, it is a biased estimator that always overestimates the parameter. Therefore, the detected change point has to be *post-estimated*, i.e., estimated after its occurrence is detected by a stopping rule.

One way of obtaining an approximately unbiased estimator of ν is to estimate the bias of $T(h)$ and subtract it from $T(h)$. According to [1], this bias, also known as *mean delay*, is asymptotically $(h + C)/K(f, g)$, as $h \to \infty$, where $K(f, g)$ is the Kullback information number, and C is a constant. In the case of sufficiently long phases before and after the change point, subtracting the estimated bias from $T(h)$ yields an approximately unbiased estimator of ν. However, in the case of frequent change points and unknown densities, no reliable estimators of C and K are available.

A *last-zero* estimator

$$\hat{\nu}_{LZ} = \sup \{k < T(h),\ W_k = 0\},$$

proposed in [20] and [26], is essentially the maximum likelihood estimator of ν, assuming a fixed-size sample rather than a sample of a random size T, which is the stopping rule. The corresponding estimator in the case of unknown densities is

$$\tilde{\nu} = \sup \left\{ k < \tilde{T}(h),\ \tilde{W}_k = 0 \right\}.$$

It can be shown that this estimator fails to satisfy an important property of *distribution consistency* ([3]).

Definition 1. *Consider a family of distributions $\mathcal{F}$ and a nonnegative function $\mathcal{D}$ on $\mathcal{F} \times \mathcal{F}$ with*

$$\mathcal{M} = \sup\{\mathcal{D}(f, g) | f, g \in \mathcal{F}\} \leq \infty.$$

Let $\mathbf{X}$ *be a fixed-size sample generated according to the multiple change-point model (2) with $\lambda \geq 1$ and $f_j \in \mathcal{F}$ for $j = 1, \ldots, \lambda$. Let $(\hat{\lambda}; \hat{\nu}_1, \ldots, \hat{\nu}_{\hat{\lambda}})$ be an estimator of $(\lambda; \nu_1, \ldots, \nu_\lambda)$. A change-point estimation scheme will be called* **distribution consistent** *(with respect to $\mathcal{D}$) if*

$$P(\hat{\lambda} = \lambda) \to 1$$

and

$$P\left(\max\left\{|\hat{\nu}_j - \nu_j|, 1 \leq j < \min(\lambda, \hat{\lambda})\right\} = 0\right) \to 1,$$

as

$$\min_{1 \leq j < \lambda} \mathcal{D}(f_j, f_{j+1}) \to \mathcal{M}.$$

This property means that all the change-point estimators converge to the corresponding parameters, as changes become more and more significant but the

sample size and all the change points remain fixed. It is implied that function $\mathcal{D}$ measures discrepancy between two distributions. Therefore, a distribution-consistent scheme estimates change points very accurately when any two consecutive phases are generated by significantly different distributions.

This property of a multiple change-point estimation scheme is desirable in all the examples of Section 1, where short phases are followed by entirely different patterns. Certainly, if the magnitude of a change is very significant, one would like to estimate the time of change very accurately, even from small samples.

Notice that for integer-valued λ and $\{\nu_j\}$, distribution consistency is equivalent to convergence in probability. It also implies that a sample with no change points ($\lambda = 0$) provides no false alarms with the probability converging to 1, and the probability of not detecting a change point in a sample with change points converges to 0.

The $\tilde{W}$-based last-zero estimator $\hat{\nu}_{LZ}$ is not distribution consistent. Indeed, any time when the last zero of $\tilde{W}$ occurs before the true change point ν, the estimator $\hat{\nu}_{LZ}$ is based on a sample from the pre-change distribution f_1 only. If f_1 remains fixed while f_2 drifts away from it so that $\mathcal{D}(f_1, f_2) \to \mathcal{M}$, the distribution of $\hat{\nu}_{NZ}$ does not change and $\hat{\nu}_{NZ}$ does not converge to ν.

For example, consider a change in the parameter of Exponential distribution from θ_0 to θ_1. In this case, the nuisance parameters are estimated by the respective sample means, and according to (3),

$$\hat{W}_n = \max_{k \le n} \{ (n-k)(\xi_{kn} - \log \xi_{kn} - 1) \},$$

where $\xi_{kn} = \bar{x}_{kn}/\bar{x}_{0k}$ and $\bar{x}_{ij} = \sum_{i+1}^{j} x_t/(j-i)$. As $\theta_1 \downarrow 0$ or $\theta_1 \uparrow \infty$ while θ_0 remains constant, $\xi_{\nu,\nu+1} - \log \xi_{\nu,\nu+1} - 1 \to \infty$ in probability. Hence,

$$P(T(h) > \nu + 1) \le P(\hat{W}_{\nu+1} \le h) \to 0, \tag{4}$$

that is, the change at ν will be detected no later than at $(\nu+1)$ with probability converging to 1. Therefore, for any $\epsilon < 1$,

$$\begin{aligned} P(|\hat{\nu}_{LZ} - \nu| < \epsilon) &= P(\hat{\nu}_{LZ} = \nu) = P(\hat{W} = 0 \cap T(h) = \nu + 1) + o(1) \\ &\to P_{\theta_0}\left(\max_{k \le \nu} (\nu - k)(\xi_{k\nu} - \log \xi_{k\nu} - 1) = 0 \right) < 1, \end{aligned}$$

as $\theta_1 \downarrow 0$ or $\theta_1 \uparrow \infty$. Thus, $\hat{\nu}_{LZ}$ (in presense of nuisance parameters) is not distribution consistent.

However, distribution-consistent schemes exist. One of them is based on the cusum stopping rule $\tilde{T}$ and the *minimum p-value* estimator

$$\hat{\nu}_{MP} = \arg \min_{1 \le k < \tilde{T}} p(k, \tilde{T}, \mathbf{X}),$$

where $p(k, \tilde{T}, \mathbf{X})$ is the p-value of the likelihood ratio test comparing subsamples $\mathbf{X_1} = (X_1, \ldots, X_k)$ and $\mathbf{X}_2 = (X_{k+1}, \ldots, X_{\tilde{T}})$.

2.3 Step 3: tests of significance

To eliminate false alarms, significance of each detected change point has to be tested. Likelihood ratio tests are easy to implement here, and significance of the detected change point is measured by the minimum p-value $p(\hat{\nu}_{MP}, \tilde{T}, \mathbf{X})$.

If the test is significant, one applies steps 1–3 to the post-change subsample $\{X_k, k > \hat{\nu}_{MP}\}$, searching for the next change point. Otherwise, we have a false alarm, and the search continues based on the initial sample, or a part of it starting after the last change point that was found significant.

2.4 The case of Gamma distributions

Gamma family is a suitable model for solution times and single-strategy use times (Fig. ?? and Fig. ??), see [3] for the results of goodness-of-fit tests.

The assumption of independence of solution times is justified by the following nonparametric test. For each of the participants who had at least two solutions, we counted the number of pairs (X_i, X_{i+1}) of consecutive solution times that are on one side of their sample median m, and the number of pairs that are on different sides. If solution times are independent in each pair, then $X_{i+1} > m$ with probability 0.5, independently of X_i. However, if X_i and X_{i+1} are positively (negatively) dependent, the probability of $(X_i - m)(X_{i+1} - m) > 0$ is greater (smaller) than 0.5. This sign test that is expressed as a simple test about the population proportion did not reject the hypothesis of independence (against a short-term dependence) with a p-value of 0.13.

For simplicity, consider a family of Gamma distributions with the same known shape parameter α and unknown scale parameter β that changes at every change point. For any two members of this family, it is natural to consider the discrepancy function

$$\mathcal{D}(f_\beta, f_{\beta^*}) = \max\{\beta/\beta^*, \beta^*/\beta\}.$$

For each $k \le \tilde{T}$, the p-value is computed as

$$p(k, \tilde{T}, R) = 1 - \left\{ \bar{B}\left(\frac{b}{b+1}, m, n\right) - \bar{B}\left(\frac{a}{a+1}, m, n\right) \right\}$$

where a and b are the only two positive roots of the equation $\Lambda(R) = \Lambda_{observed}$ for the likelihood ratio test statistic $\Lambda(R)$,

$$R = \sum_{j \le \hat{\nu}_{MP}} X_j \left(\sum_{j > \hat{\nu}_{MP}} X_j \right)^{-1},$$

and $\bar{B}$ denotes the incomplete Beta function. The statistic is computed as

$$\Lambda(R) = \left(\frac{k^k (\tilde{T}-k)^{\tilde{T}-k}}{\tilde{T}^{\tilde{T}}} (1 + R^{-1})^k (1+R)^{\tilde{T}-k} \right)^\alpha$$

and graphed on Fig. ??.

Theorem 1 *(Distribution consistency of the proposed scheme) Suppose that*

(i) *Gamma*(α_j, θ_j) *distribution changes to Gamma*$(\alpha_{j+1}, \theta_{j+1})$ *at a change point* ν_j, $j < \lambda$;

(ii) α_j *are known,* θ_j *are unknown, and* $\rho_j = \theta_{j+1}/\theta_j$;

(iii) $h \to \infty$ *and* $h/\min_j |\log \rho| \to 0$ *as* $\rho \to 0$.

Then the following probabilities converge to 1 as $\rho \to 0$*:*

(a) the probability of no false alarms, $1 - P(\tilde{T} \le \nu)$,

(b) the probability of detecting a change, $P(\tilde{T} < N \,|\, \nu \le N)$,

(c) the probability of a minimal delay, $P(\tilde{T} = \nu + 1)$,

(d) the probability that the detected change point is found significant, $P(p(\hat{\nu}_{MP}, \tilde{T}, \mathbf{X}) < \delta)$, *for any* $\delta > 0$,

(e) the probability of estimating with no error, $P(\hat{\nu}_{MP} = \nu)$.

$\hat{W}_n$ parameter at (d) variables.

Corollary 1 *The multiple change-point estimation scheme* $(\tilde{T}, \hat{\nu}_{MP})$ *is distribution-consistent.*

In fact, Theorem 1 establishes a stronger property than the distribution consistency. That is, detection of each change point requires just one post-change observation with the probability converging to 1. It is a valuable property in any problem where complicated patterns force to use minimum data to detect a change point. For the proof of Theorem 1, see [3].

2.5 Classical consistency

It is generally known ([13]) that change-point estimators are not consistent in a classical sense. It agrees with the intuition, since increasing the sample size at the expense of remote (in time) observations barely helps to estimate the change point more accurately. The smallest asymptotic error of estimation is attained by the maximum likelihood estimator $\hat{\nu}_{MLE}$. Even in the case of known distributions and exactly one change point, it has $|\hat{\nu}_{MLE} - \nu| = O_p(1)$, as $\nu, n - \nu \to \infty$.

To obtain a similar asymptotic error in presence of nuisance parameters and multiple change points, we have to modify the stopping rule,

$$\tilde{T}(\epsilon, h) = \inf \left\{ n \ : \ \max_{\epsilon \le k < n - \epsilon} \sum_{j=k+1}^{n} \log \frac{\hat{g}(X_j)}{\hat{f}(X_j)} \ge h \right\}$$

for some $\epsilon > 0$. This ensures that unknown densities are estimated from sufficiently large samples. The next theorem states that our three-step algorithm (with the threshold h being a function of the sample size N) provides the same asymptotic error of multiple change-point estimators. At the same time, the number of change points $\lambda - 1$ is estimated consistently in the classical sense.

Theorem 2 *(Sample consistency) Assume that*

(i) all the densities in (2) belong to a canonical exponential family

$$f_j(x) = f(x|\theta_j) = f(x|0)\exp\{\theta_j x - \psi(\theta_j)\};$$

(ii) $\psi''(\theta_j) > 0$ *for* $j = 1, \ldots, \lambda$;

(iii) there are no fake change points, i.e., $\theta_j = \theta_k$ *yields* $|j - k| \geq 2$;

(iv) all nuisance parameters are estimated by the method of maximum likelihood;

(v) sample size N*, threshold* h*, and the smallest distance between change points* $\Delta = \min_j \{\nu_{j+1} - \nu_j\}$ *satisfy the following conditions,*

$$\Delta \to \infty, \quad N\exp\left\{-\frac{\epsilon(N)h(\epsilon(N))}{N}\right\} \to 0, \quad \frac{h(N)}{N - \epsilon(N)} \to 0, \quad \frac{h(N)}{\Delta(N)} \to 0,$$

as $N \to \infty$.

Then
(a) $\hat{\lambda}$ *is consistent, i.e.,* $P\{\hat{\lambda} = \lambda\} \to 1$*, as* $N \to \infty$*, for any* $\lambda \geq 0$;
(b) $\max_{1 \leq j < \lambda} |\hat{\nu}_j - \nu_j| = O_p(1)$*, as* $N \to \infty$*, where* $\{\hat{\nu}_j\}$ *are change point estimates in their ascending order.*

Conditions of this theorem are trivially satisfied if, say, $\epsilon(N) = \epsilon N$, $\Delta(N) = \Delta N$, and $h(N)$ is any function whose rate of growth is between $\log(N)$ and N.

3 Applications and practical results

We used the described algorithm to identify global changes of climate, spikes in hourly electricity prices, and different phases in development, learning, and problem solving processes.

Analysis of Central England Temperatures data ([19],[21]) shows significant changes of climate around 1730, 1830, and in the 1940s. The former two may be attributed to the Little Ice Age period whereas the latter is likely to be related to a greenhouse effect ([1]).

Analysis of detrended electricity prices (Fig. ??) allows to separate spikes and to fit a suitable Markovian model with transitions from the regular state to

the spike state and vice versa. Then, an appropriate ARMA model is fit to the transformed interspike prices, whereas spikes are modeled by a compound log-normal distribution ([4]). Such a stochastic model is necessary for the prediction of prices, evaluation of futures and forward options, etc.

Analysis of the solution times and single-strategy use times shows a number of different phases in microdevelopmental processes. Comparison of different phases discovers that some earlier patterns are repeated later. A clustering algorithm is then used to match similar patterns and to identify different "types of behavior" ([2]).

Acknowledgements. We thank Professors Yoel Haitovsky, Udi Makov, and Ya'acov Ritov for the very well organized conference and their hospitality in Israel. Research of both authors is supported by the U.S. National Science Foundation, grant # 9818959.

References

[1] M. Baron. Nonparametric adaptive change-point estimation and on-line detection. *Sequential Analysis*, 19:1–23, 2000.

[2] M. Baron and N. Granott. Estimation of frequent change points. *In preparation*, 2001.

[3] M. Baron and N. Granott. Small sample change-point analysis with applications to problem solving. *Under review*, 2001.

[4] M. Baron, M. Rosenberg, and N. Sidorenko. Automatic spike detection for the modeling and prediction of electricity pricing. *Energy & Power Risk Management*, 2001.

[5] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. PTR Prentice-Hall, Inc., 1993.

[6] P. K. Bhattacharya. Some aspects of change-point analysis. In *Change-Point Problems*, pages 28–56, IMS Lecture Notes - Monograph Series (Volume 23), 1995.

[7] B. E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, The Netherlands, 1993.

[8] E. Carlstein. Nonparametric estimation of a change-point. *Ann. Statist.*, 16:188–197, 1988.

[9] H. Chernoff and S. Zacks. Estimating the current mean of a normal distribution which is subject to changes in time. *Ann.Math.Statist.*, 35:999–1028, 1964.

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

[11] R. Ethier and G. Dorris. Don't ignore the spikes. *Energy & Power Risk Management*, July/August, 1999.

[12] Y.-X Fu, R. N. Curnow. Maximum likelihood estimation of multiple change points. *Biometrika*, 77:563–573, 1990.

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

[14] R. A. Khan. A sequential detection procedure and the related cusum procedure. *Sankhya*, 40:146–162, 1979.

[15] T. L. Lai. Sequential changepoint detection in quality control and dynamical systems. *J. R. Statist. Soc. B*, 57:613–658, 1995.

[16] C.-B. Lee. Estimating the number of change points in a sequence of independent normal random variables. *Statist. Probab. Letters*, 25:241–248, 1995.

[17] C.-B. Lee. Nonparametric multiple change-point estimators. *Statist. Probab. Letters*, 27:295–304, 1996.

[18] G. Lorden. Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, 42:1897–1908, 1971.

[19] G. Manley. Central England Temperatures: monthly means 1659 to 1973. *Q.J.R. Meteorol. Soc.*, 100:389–405, 1974.

[20] D. C. Montgomery. *Introduction to Statistical Quality Control, Third Ed.* Wiley, New York, 1997.

[21] D. E. Parker, T. P. Legg, and C. K. Folland. A new daily central england temperature series, 1772-1991. *Int. J. Clim.*, 12:317–342, 1992.

[22] M. Pollak. Optimal detection of a change in distribution. *Ann. Statist.*, 13:206–227, 1985.

[23] M. Pollak and D. Siegmund. Sequential detection of a change in a normal mean when the initial value is unknown. *Ann. Statist.*, 15:749–779, 1987.

[24] Y. Ritov. Decision theoretic optimality of CUSUM procedure. *Ann. Statist.*, 18:1464–1469, 1990.

[25] A. N. Shiryaev. On optimal method in earliest detection problems. *Theory Probab. Appl.*, 8:26–51, 1963.

[26] M. S. Srivastava, Y. Wu. Quazi-stationary biases of change point and change magnitude estimation after sequential cusum test. *Sequential Analysis*, 18:203–216, 1999.

[27] L. Ju. Vostrikova. Detecting "disorder" in multidimensional random processes. *Sov. Math. Dokl.*, 24:55–59, 1981.

[28] S. Zacks. Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures in testing and estimation. In *Recent Advances in Statistics*, pages 245–269, Academic Press, 1983.