ON DETECTING JUMPS IN TIME SERIES - NONPARAMETRIC SETTING

Mirek Pawlak
University of Manitoba, Canada

Ewaryst Rafajłowicz
Technical University of Wrocław, Poland

Ansgar Steland[1]
Ruhr-Universität Bochum, Germany

ABSTRACT. Motivated by applications in statistical quality control and signal analysis, we propose a sequential detection procedure which is designed to detect structural changes, in particular jumps, immediately. This is achieved by modifying a median filter by appropriate kernel-based jump preserving weights (shrinking) and a clipping mechanism. We aim at both robustness and immediate detection of jumps. Whereas the median approach ensures robust smooths when there are no jumps, the modification ensure immediate reaction to jumps. For general clipping location estimators we show that the procedure can detect jumps of certain heights with no delay, even when applied to Banach space valued data. For shrinking medians we provide an asymptotic upper bound for the normed delay. The finite sample properties are studied by simulations which show that our proposal outperforms classical procedures in certain respects.

Keywords: Edge Detection, Nonparametric Estimation, Quality Control, Statistical Process Control.
MSC: 62L15, 62L10, 62G08.

Date: Sept 2003.

[1]Address of correspondence: Ansgar Steland, Ruhr-Universität Bochum, Fakultät für Mathematik, Mathematik 3 NA 3/71, Universitätsstr. 150, D-44780 Bochum, Germany.

1. Introduction

Our aim is to study a class of sequential detection rules. The basic situation is as follows. We are given a sequential stream of observations $Y_1, Y_2, \ldots$ with associated time stamps $t_1, t_2, \ldots$. The observations represent certain (univariate or multivariate) quality characteristics or a sequence of signals, e.g., FFT spectra obtained by analyzing frames of an audio signal. In many applications one is interested in on-line monitoring of such sequences of observations meaning that one wants to get a signal, if there is some departure from normal behavior. In statistical terms it is assumed that the sequence $\{Y_n\}$ is distributed according to a known in-control model corresponding to a certain null hypothesis about the process and therefore about the reality. At each time-point t_n a statistical decision procedure is applied to the available data set $Y_1, \ldots, Y_n$ in order to decide whether we still should rely on the in-control model or whether there is strong enough evidence to reject this model.

In this paper we study a general detection rule which is especially designed to detect sufficiently high jumps with no delay. This means, our procedure can guarantee that the location of the jump is exactly reproduced. Further, the method can be applied to various fields, since it can deal with univariate, multivariate, and even function-valued data.

To motivate our approach, let us briefly consider the following application in some detail. In Statistical Quality Control one observes a quality characteristic, Y_n, at each time point t_n. In many cases the data $\{(Y_n, t_n)\}$ is obtained by sampling the underlying continuous-time process at discrete ordered time-points $\{t_n\}$. It is assumed that $\{Y_n\}$ forms a sequence of identically and symmetrically distributed random variables with common median m, as long as the production process is in-control. If there is a failure, certain characteristics of the process may change, and based on statistical estimators of these quantities we may infer that a change has occurred. For simplicity we confine ourselves to the case that the median changes, since in general one can transform the data to ensure that structural changes affect the location of the process. Detecting jumps immediately and in a robust way is often crucial to analyze the cause of the failure. This in particular applies for complex production processes where a large number of systems interact. In this case a severe failure is often the result of a cascade of small failures at different time points. Such a complex failure can also change the distribution consideraby. Thus, robustness is a concern.

The best known classical procedures to detect change-points sequentially are the likelihood motivated CUSUM chart, which dates back to Page (1954, 1955), the Shiryaev-Roberts procedure, independently proposed by by Girshick and Rubin (1952) and Shiryaev (1950), and the EWMA chart due to Roberts (1959), which employs the L_2-optimal predictor of the integrated moving average process of order 1. For results on the (asymptotic) optimality properties of the CUSUM and Shiryaev-Roberts procedure we refer to Pollak (1985), Moustakides (1986), Ritov (1990), and Yakir (1996, 1997). Since these procedures are motivated by maximum likelihood and Bayesian approaches, one has to know the in-control and out-of-control distributions to calculate the relevant statistics. Both the CUSUM and EWMA procedures have been also extended and intensively studied for dependent processes, too. The basic idea is to use these schemes as motivated by a certain parametric model and to modify the procedure to take account of dependencies or different distributions. For details of that approach we refer to Vasilopoulos and Stamboulis (1978), Schmid and Schoene (1997), and the references given there.

From a nonparametric viewpoint a natural candidate procedure is to use a nonparametric estimator of the process mean and to compare it with some critical value. Smoothing estimators which estimate smooth functions consistently were intensively studied in the classical fixed-sample design. Sequential procedures based on related kernel smoothers and optimal kernels have been studied by Schmid and Steland (2000) and Steland (2003a, 2003c). Asymptotic distribution theory for dependent time series can be found in Steland (2003b). For an application to sequential control of credit risk management see Steland (2002b). Since the procedures studied there are based on weighted averages, they are inherently not robust. Furthermore, classic smoothers then to smooth large jumps. Therefore, in this paper we investigate the performance of a jump preserving procedure based on the median.

Whereas classical nonparametric smoothers only use horizontal smoothing, jump-preserving estimators also employ a vertical smoothing scheme. It is that property which enables jump-preserving estimators to reproduce jumps more accurately than other approaches. The basic idea of this approach was developed for image processing purposes and is called sigma filter (Lee, 1983). Related techniques, in particular robust approaches, have been studied by Chiu et. al. (1998) with an emphasis towards image processing, and by Rue et al. (2002) using local linear M-smoothers. Rafajłowicz (1996) proposed their application to sequential monitoring. Pawlak and Rafajłowicz (2000, 2001c) studied the more general framework of vertical regression. Limit theorems for the normed delay of stopping rules relying on classical kernel smoothers can be found in Brodsky and Darkhovsky (1993, 2000). An extension to the sigma filter for i.i.d. data was given by Steland (2002a). The application of a median-based clipping estimator was proposed by Kryzak, Rafajłowicz and Skubalska-Rafajłowicz (2001).

2. Sequential detection procedure

The proposed method of detection is probably best understood in the context of univariate observations. After describing the change-point model, we give an introduction to jump-preserving median estimation. The extension to higher dimensional data is discussed in the next section.

2.1. **Model.** Let $Y = Y(t)$ be a generic univariate observation with associated time t. We assume that

$$Y(t) = m(t) + \epsilon_t$$

where ϵ_t denotes an error term distributed according to a symmetric density f_ϵ with median 0. Consequently, $m(t)$ stands for median of $Y(t)$. Further, we will denote the density of Y at time t by $f(y;t)$.

For a given stream $\{(Y_i, t_i)\}$ of observations obtained by observing $Y(t)$ at the ordered non-stochastic time points $t_1 < t_2 < \dots$, we assume that

$$\text{Med}(Y_i) = 0, \qquad t_i < t_q \qquad \textit{(in-control model)}$$

and

$$\text{Med}(Y_i) = m(t_i) > 0, \qquad t_i \geq t_q \qquad \textit{(out-of-control model).}$$

More precisely, for our results we assume that

$$\text{Med}\, Y(t) = m(t) \geq B > 0$$

if the process runs out-of-control, i.e, $t \geq t_q$. Here t_q denotes the change-point and q its index. Of course, one may also consider negative shifts, but for simplicity of presentation we shall concentrate on positive ones.

2.2. **Jump-preserving estimation.** Before presenting the method itself, we shall provide a brief introduction to the statistical reasoning leading to our approach.

Recall that the median is a minimizer of the expected absolute deviation,

$$m(t) = \text{argmin}_{m \in \mathbb{R}} \int |y - m| f(y;t)\, dy. \tag{1}$$

The basic idea of our approach is to introduce a weighting mechanism in this objective function which ensures that the minimizer is still given by the median $m(t)$, and which simultaneously guarantees that observations far away from $m(t)$ are downweighted. Define

$$k_M(y;t) = k\left(\frac{y - m(t)}{M}\right)$$

where k is a symmetric probability density with center of symmetry equal to 0 and M denotes a positive parameter. Instead of (1) we shall now study

$$m^*(t) = \text{argmin}_{m\in\mathbb{R}} \int k_M(y;t)|y-m|f(y;t)\,dy. \tag{2}$$

Of course, our aim is to verify that $m^*(t) = m(t)$. This fact can be easily seen by a change of measure. Observe that

$$\begin{aligned} E[k_M(y;t)|Y-m|] &= \int k_M(y;t)|y-m|f(y;t)\,dy \\ &= c_{k_M;t}^{-1}\int |y-m|f_{k_M}(y;t)\,dy \end{aligned} \tag{3}$$

where the transformed density $f_{k_M}(y;t)$ is given by

$$f_{k_M}(y;t) = \frac{f(y;t)k_M(y;t)}{c_{k_M;t}}, \qquad c_{k_M;t} = \int f(y;t)k_M(y;t)dy.$$

It is instructive to note that for kernels k with support $[-1,1]$ the transformed density has support $[m(t)-M, m(t)+M]$. Moreover, for the uniform kernel given by $k(z) = 1/2$ if $|z| \le 1$ and $k(z) = 0$ otherwise, the transformed density is simply obtained by truncating the distribution and renormalizing. Equation (3) shows that if Y^* is distributed according to $f_{k_M}(y;t)$, we may write

$$m^*(t) = \text{Med}(Y^*;t).$$

As a consequence, a minimizer of the weighted version (2) is given by the median of an observation Y^* which is distributed according to the transformed density which has support $[m(t)-M, m(t)+M]$.

To see that the median of Y^* is equal to $m(t)$, note that the transformed density is obtained by a translation $y \mapsto y - m(t)$ of the density

$$\varphi_{k_M}(z;t) = \frac{f_\epsilon(z)k(z/M)}{c_{k_M;t}},$$

which is symmetric around 0, since the error density f_ϵ and the kernel k have this property.

2.3. **Sampling from f_{k_M} and a data transformation.** Let us briefly discuss a heuristic but not rigoruous reasoning of an approach to obtain a sample whose distribution is approximately given by $f_{k_M}(y;t)$, and a data transformation to obtain a sample catching the location. Our remarks only aim at providing further motivation. The rigorous results presented below are not affected by these rough ideas.

Let $-\infty < g_1 < \cdots < g_R < \infty$ be equidistant points with $\Delta = g_{j+1} - g_j$. Choose $\xi_j \in (g_j, g_{j+1}]$. Recall that the histogram approximates the underlying density, i.e., $f_j/\Delta \approx$

$f(\xi_j)$, if f_j is the proportion of the $Y_i's$ in $(g_j, g_{j+1}]$. Suppose $Y_1^*, \dots, Y_L^*$ is a sample with corresponding proportions

$$f_j^* = \frac{k(\xi_j/M) f_j}{\sum_{l=1}^L k(\xi_l/M) f_l}, \; j = 1, \dots, R-1.$$

Note that such a sample can be approximately constructed if we use $\lfloor f_j^* f_j^{-1} L \rfloor$ copies of each observation in $(g_j, g_{j+1}]$. Here $\lfloor x \rfloor$ denotes the greatest integer less or equal than x, $x \in \mathbb{R}$. We have

$$\begin{aligned} \frac{f_j^*}{\Delta} &= \frac{k(\xi_j/M) f_j/\Delta}{\sum_l k(\xi_l) f_l} \\ &\approx \frac{k(\xi_j/M) f(\xi_j)}{\sum_l k(\xi_j/M) f(\xi_l)\Delta} \\ &\approx \frac{k(\xi_j/M) f(\xi_j)}{\int k(x/M) f(x)\, dx} = f_{k_M}(\xi_j). \end{aligned}$$

To motivate a data transformation of $Y_1, \dots, Y_n$ related to the first moment of f_{k_M}, suppose now that $f_j = 1/n$ for all j, meaning that f_j/Δ yields an approximation of $f(Y_{(j)})$, where $Y_{(1)}, \dots, Y_{(n)}$ denotes the order statistic. Then both sides of

$$f_j^*/\Delta \approx \frac{k(Y_{(j)}/M) f(Y_{(j)})}{\sum_l k(Y_{(j)}/M) f(Y_{(j)})\Delta}$$

yield an approximation of $f_{k_M}(y)$ at $y = Y_{(j)}$. Note that the f_j^* define a discrete distribution which puts mass $k(Y_i/M)/\sum_j k(Y_j/M)$ on the point Y_i, $i = 1, \dots, n$. The moments of this distribution are given by

$$m_r = \frac{\sum_{i=1}^n k(Y_i/M) Y_i^r}{\sum_{i=1}^n k(Y_i/M)}, \qquad r \in \mathbb{N}.$$

Looking at the numerator and neglecting the denominator, we propose to base inference on the transformed observation

$$k([Y - m(t)]/M) Y \tag{4}$$

The median of the random variable (4) is

$$\begin{aligned} \mathrm{Med}[k([Y - m(t)]/M) Y] &= \mathrm{Med}[k(\epsilon/M) m(t)] + \mathrm{Med}[k(\epsilon/M)\epsilon] \\ &= m(t)\, \mathrm{Med}[k(\epsilon/M)], \end{aligned}$$

since $\epsilon \stackrel{d}{=} -\epsilon$ and $k(-x) = k(x)$ for all $x \in \mathbb{R}$, which implies $k(\epsilon/M)\epsilon \stackrel{d}{=} k(-\epsilon/M)(-\epsilon) = -k(\epsilon/M)\epsilon$, i.e., $\mathrm{Med}[k(\epsilon/M)\epsilon] = 0$. Consequently, the median of the transformed observations (4) is proportional to $m(t)$.

To obtain data taking on extreme values if $|m(t_n)|$ is large, we plug in Y_n as an estimate for $m(t_n)$. Hence we base our procedures on the transformed sample

$$Z_i = k([Y_i - Y_n]/M)Y_i, \qquad 1 \leq i \leq n, \; n \in \mathbb{N}.$$

Note that observations far away from the current observation are shrunken towards 0. Further, if Y_i is symmetrically distributed around 0 (in-control model), then $Y_i \stackrel{d}{=} -Y_i$, and therefore

$$-k([Y_i - Y_n]/M)Y_i \stackrel{d}{=} k([Y_i - Y_n]/M)Y_i.$$

Consequently, the median of these transformed quantities is 0.

2.4. **The detection procedure.** To obtain a robust and smooth estimate when the process is in control, we propose to calculate the empirical (clipping) median of the most recent, say, h observations of the sample $Y_1, \ldots, Y_n$ available at time t_n. Here the clipping median is defined as

$$\widehat{m}_{nh} = \text{ClipMed}_{n-h+1 \leq i \leq n} \left\{ k\left(\frac{Y_i - Y_n}{M} \right) Y_i \right\},$$

where ClipMed stands for the empirical median calculated from all Z_i's with corresponding Y_i satisfying $|Y_i - Y_n| \leq M$. Note that observations with $|Y_i - Y_n| > M$ are excluded from the calculation of the median. Taking the median ensures a certain degree of robustness of the estimator if there are no jumps, whereas the clipping property ensures that the estimator is able to react immediately when there is a level shift or strong trend in the data, for the following reasons. Both properties are controlled by h (degree of smoothing) and M (sensitivity with respect to jumps). If M is not too small, on average only a few observations will be excluded reducing efficiency only slightly. But if there is a (large) jump, Y_n is expected to be large, and thus the neighorbood defined by $\mathcal{N}(Y_n) = \{i : |Y_i - Y_n| \leq M\}$ will shrink substantially. This has the effect that, firstly, the median is calculated from a small sample mainly consisting of observations after the change point, and, secondly, the estimator no longer smoothes the data. Note that if $\mathcal{N}(Y_n) = \{n\}$, which may happen if Y_n is an extreme outlier or the first observation with a mean differing considerably from the in-control mean, then the clipping median interpolates, i.e., $\text{ClipMed}_{n-h+1 \leq i \leq n} = Y_n$. If the process mean is constant after the change point, the estimator will stabilize again, since $\mathcal{N}(Y_n)$ will tend to increase successively, and continues to smooth the data in a robust way.

Noting that we may assume that the in-control median is 0, since otherwise we may replace the Y_is by $Y_i - \mu_0$ where μ_0 is the in-control median, we decide in favor of a jump at time n if

$$\widehat{m}_{nh} < -c, \quad \widehat{m}_{nh} > c, \quad \text{or} \quad |\widehat{m}_{nh}| > c$$

for some pre-specified threshold (critical value) $c > 0$, depending on whether one aims at detecting positive, negative, or arbitrary jumps. The corresponding stopping time for the two-sided case is therefore defined as

$$N_h = N_h(c) = \inf\{n \in \mathbb{N} : |\widehat{m}_{nh}| > c\}.$$

2.5. **Choice of the threshold.** Concerning the choice of the threshold c there are two basic approaches. A popular criterion is to choose c such that the average in-control run length attains a pre-specified value, say, ξ. The average in-control run length is defined by

$$\mathrm{ARL}_0(c; h) = E[N_h(c)] = \sum_{k \in \mathbb{N}} k P_0(N_h(c) = k)$$

The value c can be approximated numerically by estimating $\mathrm{ARL}_0(c; h)$ by simulation. For better comparison we used this approach in our simulation study.

However, this requires knowledge of the in-control distribution, which is often unknown. Further, for some applications it is more important to detect jumps of a certain magnitude with a high certainty than knowing that the stopping time has a certain average run length. Thus, one can also choose c depending on the size of a jump we want to detect with high probability. This approach is also motivated by the no-delay property discussed below.

3. Zero-delay property

In this section we discuss the zero-delay property of the proposed sequential detection procedure. We provide sufficient conditions on the design of the detection procedure such that a jump of height B is detected with no delay. It turns out that this property holds true for univariate, multivariate, and function-valued observations. Further, no assumptions on the dependence structure are necessary. This means, the result is valid for any dependent stochastic process. When concerned with multivariate observations, e.g., vectors, a *jump* of height B is a vector B defining the jump height for each coordinate. It is clear that we need an appropriate framework which can deal with the various cases. Basically, we have certain objects y representing possible observations, addition, subtraction of these objects is required and should have a reasonable interpretation. Of course, multiplication with real numbers is necessary, and finally, an order relation to compare objects. Therefore, we shall assume that we are given Banach space-valued observations where the Banach space is equipped with an order relation.

To this end let $(\mathcal{B}, \| \circ \|)$ be a Banach space with norm $\| \circ \|$. Let P denote a probability measure defined on a sigma algebra $\mathcal{F}$ of measurable events $A \subset \mathcal{B}$. If the dimension of $\mathcal{B}$

is finite, all norms are equivalent, but the choice of the norm is important for applications, since it defines our understanding of neighborhoods.

We shall further assume that there is an order relation $\leq$ given, i.e., a subset $\mathcal{R}_{\leq} \subset \mathcal{B} \times \mathcal{B}$ such that for all $x, y, z \in \mathcal{B}$ the following three properties are satisfied:

(i) $(x, y) \in \mathcal{R}_{\leq}$ or $(y, x) \in \mathcal{R}_{\leq}$.
(ii) $(x, x) \in \mathcal{R}_{\leq}$.
(iii) If $(x, y) \in \mathcal{R}_{\leq}$ and $(y, z) \in \mathcal{R}_{\leq}$, then $(x, z) \in \mathcal{R}_{\leq}$.

As usual, we write $x \leq y$ if $(x, y) \in \mathcal{R}_{\leq}$.

Let $Y_1, Y_2, \ldots$ be a sequence of independent and identically distributed $\mathcal{B}$-valued random elements. We shall now generalize the detection rule proposed in the previous section to the current more general framework. Note that the median-based estimator proposed in the previous section is well-defined if $\mathcal{B} = \mathbb{R}$ and $\| \circ \| = | \circ |$. For the multivariate case, $\mathcal{B} = R^k$ and $\| \circ \|$ being an arbitrary vector norm, the multivariate median is not uniquely defined. However, the zero-delay property does in fact not depend on the concrete choice of a location estimator, as long as the estimator satisfies a rather mild regularity assumptions, which is satisfied by many location estimators. Therefore we assume that

$$\widehat{m}_{nh} = \text{LocEst}_{i \in \{n-h+1, \ldots, n : |Y_i - Y_n| \leq M\}} \left\{ k\left(\frac{Y_i - Y_n}{M}\right) Y_i \right\}, \tag{5}$$

where LocEst stands for a clipping location estimator. This means, $\widehat{m}_{nh}$ is a statistic with values in $\mathcal{B}$ depending only on the random elements $Y_1, \ldots, Y_n$ up to time t_n, such that only random elements Y_i are used if their distance to the current observation Y_n does not exceed M. More precisely, we assume that

$$\widehat{m}_{nh} \text{ is a } \sigma(\mathcal{Y}_{n-h+1}^n)\text{-measurable statistic} \tag{6}$$

where

$$\mathcal{Y}_l^n = \{Y_i : l \leq i \leq n, |Y_i - Y_n| \leq M\}$$

with

$$\min \|\mathcal{Y}_{n-h+1}^n\| \leq \|\widehat{m}_n\| \leq \max \|\mathcal{Y}_{n-h+1}^n\|. \tag{7}$$

Here we use the notation

$$\min \|\mathcal{A}\| = \min\{\|a\| : a \in \mathcal{A}\}.$$

$\max \|\mathcal{A}\|$ is defined analogously. A signal to indicate evidence that the process is no longer in control is given when

$$\|\widehat{m}_n\| > c \tag{8}$$

where c is a pre-specified critical value (threshold).

We are now in a position to formulate the result on zero-delay detection in a rigorous fashion.

Theorem 3.1. *Assume (6), (7). Further, suppose that the process mean $m(t)$ satisfies*

$$m(t) = \begin{cases} 0, & t < t_q \text{ (in-control model)} \\ B, & t \geq t_q \text{ (out-of-control model)} \end{cases}$$

For any stationary stochastic process $\{\epsilon_n : n \in \mathbb{Z}\}$ in discrete time with

$$P[-A \leq \epsilon_n \leq A] = 1$$

holds true for some constant $A \in \mathcal{B}$, the detection rule (8) with $\widehat{m}_{nh}$ defined by (5) detects the jump B with zero delay, with probability 1, if

$$\|B - A\| - M > c.$$

Proof. Using $-A \leq \epsilon_n \leq A$, for all n, we may argue as follows. For all $n < q$ we have

$$Y_n = m(t_n) + \epsilon_n \leq A,$$

whereas for $n \geq q$

$$Y_n = m(t_n) + \epsilon_n \geq B - A.$$

Further, by definition of $\widehat{m}_n$ and $\mathcal{Y}_{n-h+1}^n$, we have for all $n \geq q$

$$\begin{aligned} \|\widehat{m}_{nh}\| &\geq \min \|\mathcal{Y}_{n-h+1}^n\| = \min_{i:|Y_i - Y_n| \leq M} |Y_i| \\ &\geq \|Y_n\| - M \geq \|B - A\| - M. \end{aligned}$$

Since we have zero delay iff. $\widehat{m}_{qh} > c$, a sufficient condition is

$$\|B - A\| - M > c.$$

□

4. The asymptotic delay for shrinking medians

In the previous section it was shown that clipping location estimators can detect jumps of certain heights with no delay for all h. This is essentially due to clipping and boundedness of the error terms. We will now omit clipping, but still shrink the data towards 0 using a kernel k. To avoid that part of the sample collapses to 0, we assume that k attaches a minimal weight, i.e.,

$$0 < k_{\min} \leq k(z) \leq k_{\max} < \infty, \text{ and } k(z) = k_{\min}, \text{ if } |z| > M. \tag{9}$$

We also assumed that $k(-z) = k(z)$ for all $z \in \mathbb{R}$ and $k(|z_1|) \geq k(|z_2|)$ for $|z_1| \leq |z_2|$. We study the asymptotic behavior in terms of the point-wise false-alarm rate and show that, roughly speaking, for large enough h the delay, when expressed as a percentage of the bandwidth h, is not larger than $(1/2 + \varepsilon)$. The results are verified for independent univariate data having a density with bounded support. The results of this section are asymptotic, but the case of small h is studied to some extent via simulations in the next section.

To this end consider the stopping rule which gives a signal if

$$\widehat{m}_{nh} > c \tag{10}$$

for a pre-specified threshold c, where

$$\widehat{m}_{nh} = \mathrm{Med}_{n-h+1 \leq i \leq n} \left\{ k\left(\frac{Y_i - Y_n}{M} \right) Y_i \right\}.$$

The associated stopping time is again given by

$$N_h = \inf\{n \in \mathbb{N} : \widehat{m}_{nh} > c\}.$$

Recall that this procedure uses the most recent h observations. Define also the related quantities delay and normed delay by

$$D_h = \max\{0, N_h - t_q\} \qquad \text{and} \qquad \rho_h = \frac{D_h}{h},$$

respectively. To study asymptotic properties of the delay it is common to consider the normed delay given by ρ_h, which measures the delay expressed as a percentage of the effective sample size h.

We will need the following lemma.

Lemma 4.1. *For $y \in \mathbb{R}$ and $M > 0$ define $Z_i(y; M) = k([Y_i - y]/M)Y_i$, $i = 1, \ldots, n$. Denote the d.f. of $Z_i(y; M)$ by $F_{Z_1(y;M)}$. We have for all $y \in \mathbb{R}$ and $M > 0$:*

(i) $F(x/k_{\max}) \leq F_{Z_1(y;M)}(x) \leq F(x/k_{\min})$, $x \geq 0$.
(ii) $F(x/k_{\min}) \leq F_{Z_1(y;M)}(x) \leq F(x/k_{\max})$, $x < 0$.

Proof. If $x \geq 0$, then $Z_1(y; M) \geq 0$ implies $Y_1 \geq 0$. Thus,

$$\{k_{\min} Y_1 \geq x\} \subset \{Z_1(y; M) \geq x\} \subset \{k_{\max} Y_1 \geq x\},$$

which verifies (i). Analogously, if $x < 0$ we have

$$\{k_{\min} Y_1 \leq x\} \subset \{Z_1(y; M) \leq x\} \subset \{k_{\max} Y_1 \leq x\}$$

verfiying (ii). □

The following results asserts that the in-control probability that a signal is given, the false-alarm rate, tends to zero, in probability, as $h \to \infty$.

Theorem 4.1. (In-control behavior). *Let $\{Y_n : n \in \mathbb{Z}\}$ be i.i.d. with common density function f satisfying $f(x) = f(-x)$ for all $x \in \mathbb{R}$ and $f(0) > 0$. Then for each $x > 0$*

$$P[\widehat{m}_{nh} > x] = o_P(1),$$

as $h \to \infty$.

Proof. By conditioning on Y_n we have to analyze

$$\int P[\mathrm{Med}_{n-h+1 \le i \le n}\{Z_i(y; M)\} > x] dF(y)$$

where for $i = n-h+1, \dots, n$

$$Z_i(y; M) = k([Y_i - y]/M) Y_i.$$

Note that sgn $Z_i(y; M) =$ sgn Y_i. Hence $P[Z_i(y; M) \le 0] = P[Y_i \le 0] = 1/2$ yielding

$$\mathrm{Med}[Z_1(y; M)] = 0, \qquad \forall y \in \mathbb{R},\ \forall M > 0.$$

Furthermore, by Lemma 4.1 we may estimate the density $f_{Z_1(y;M)}$ of the variables $Z_i(y; M)$ by

$$\begin{aligned} f_{Z_1(y;M)}(0) &= \lim_{\varepsilon \downarrow 0} \frac{F_{Z_1(y;M)}(\varepsilon) - F_{Z_1(y;M)}(-\varepsilon)}{2\varepsilon} \\ &\ge \lim_{\varepsilon \downarrow 0} \frac{1}{k_{\max}} \frac{F(\varepsilon/k_{\max}) - F(\varepsilon/k_{\max})}{2\varepsilon/k_{\max}} \\ &= k_{\max}^{-1} f(0). \end{aligned}$$

Thus, we obtain

$$0 < f(0)/k_{\max} \le \inf_{y \in \mathbb{R}, M > 0} f_{Z_1(y;M)}(0).$$

We shall use the Bahadur-type representation of the median, see Bahadur (1966), Kiefer (1967), Serfling (1980, ch. 2.5.2), and Hesse (1990). It is known that

$$\mathrm{Med}_{n-h+1 \le i \le n}\{Z_i(y; M)\} = \frac{1}{h} \sum_{i=n-h+1}^{n} \eta_i(y; M) + R_n, \tag{11}$$

as $h \to \infty$, where

$$\eta_i(y; M) = \frac{\mathbf{1}(Z_i(y; M) \le 0) - 1/2}{f_{Z_1(y;M)}(0)},$$

$n - h + 1 \le i \le n$, and the remainder term satisfies

$$R_n = O(h^{-3/4} (\log h)^{1/2} (\log \log h)^{1/4}).$$

with probability 1, more precisely, Kiefer shows that

$$\limsup_{n\to\infty} \frac{n^{3/4} R_n}{(\log\log n)^{3/4}} = \frac{2^{5/4}[p(1-p)]^{1/4}}{3^{3/4}},$$

where in our case $p = 1/2$. Hence, the approximation (11) is uniform in $y \in \mathbb{R}$ and $M > 0$. Further, the random variables $\{\eta_i(y; M)\}$ are i.i.d. with

$$|\eta_i(y; M)| \leq M = \sup_{y\in\mathbb{R}, M>0} f^{-1}_{Z_1(y;M)}(0).$$

Consequently, Bernstein's inequality gives

$$P[\sum_{i=n-h+1}^{n} \eta_i(y; M) > x] \leq 2\exp\left(-\frac{1}{2}\frac{(hx)^2}{v + Mx/3}\right),$$

where

$$v > \sup_y \mathrm{Var}\,(\sum_{i=n-h+1}^{n} \eta_i) = h(1/4)/\inf_{y\in\mathbb{R}, M>0} f_{Z_1(y;M)}(0).$$

□

The next result considers the (asymptotic) delay of the rule (10). Due to the robustness of the median, without clipping we can not expect to obtain a no-delay property. However, the delay is not larger than $(1/2 + \varepsilon)h$ for arbitrary $\varepsilon > 0$, if the bandwidth is sufficiently large and the jump is large enough. Interestingly, it turns out that in order to detect small jumps with minimal delay, the minimal weight $k_{\min}$ attached by the kernel $k_{\min}$ should be not too small.

Theorem 4.2. (Out-of-control behavior). *Assume both the kernel k satisfies condition (9) and the process mean $m(t)$ satisfies*

$$m(t) = \begin{cases} 0, & t < t_q \text{ (in-control model)} \\ B, & t \geq t_q \text{ (out-of-control model)} \end{cases}$$

Assume $\{\epsilon_n\}$ are i.i.d. with support $[-A, A]$ for some $A > 0$, $f(x) = f(-x)$ for all $x \in \mathbb{R}$, and $f(0) > 0$. If

$$B \geq 2A + M, \quad B > \max\{A(k_{\max} + k_{\min})/k_{\min}, [c + k_{\max}]A/k_{\min}\},$$

then for each $\varepsilon > 0$ there exists some $h_0 > 0$ such that for $h \geq h_0$ the normed delay, ρ_h, of the procedure (10) satisfies

$$P[\rho_h > 1/2 + \varepsilon] = 0.$$

Proof. Let $\varepsilon > 0$. We have

$$P[\rho_h > 1/2 + \varepsilon] = P[N_h > t_q + (1/2 + \varepsilon)h] \leq P[\widehat{m}_{l(h),h} \leq c].$$

where $l(h) = t_q + \lfloor (1/2 + \varepsilon)h \rfloor$. Clearly, if $h_0 = 1/(2\varepsilon)$ we have $(1/2+\varepsilon)h \geq (h+1)/2$ for all $h \geq h_0$. Put $Z_i = k([Y_i - Y_{l(h)}]/M)Y_i$ for all i. If $i < t_q$ we have $Z_i = k([\epsilon_i - (\epsilon_{l(h)} + B)]/M)\epsilon_i$. Since $B \geq 2A + M$ and $\epsilon_{l(h)} + B \in [B - A, B + A]$, $\epsilon_i \leq A \leq B - M - A \leq \epsilon_{l(h)} + B - M$, i.e., $Z_i = k_{\min}\epsilon_i$ by definition of the kernel k. Consequently,

$$Z_i \in [-k_{\min}A, k_{\min}A] \qquad \text{if } i < t_q.$$

For $i \geq t_q$ we have $Z_i = k([\epsilon_i - \epsilon_{l(h)}]/M)\epsilon_i$ yielding

$$Z_i \in [k_{\min}B - k_{\max}A, k_{\max}(A + B)] \qquad \text{if } i \geq t_q.$$

Since the breakpoint of the median is 1/2,

$$\text{Med}_{i=l(h)-h+1,\ldots,l(h)}\{Z_i\} \geq k_{min}B - k_{\max}A,$$

if $k_{\min}A < k_{\min}B - k_{\max}A$ or, equivalently, $B > A(k_{\max} + k_{\min})/k_{\min}$. Thus, if additionally $k_{\min}B - k_{\max}A > c$, or, equivalently, $B > [c + k_{\max}]A/k_{\min}$, $\{\widehat{m}_{l(h),h} \leq c\} = \emptyset$.

□

5. Simulations

To shed some light onto the performance properties of the methods studied in this paper we performed a simulation study. The primary aims were to analyze (i) whether the procedures are able to detect change-points with high probability when confronted with error distributions with unbounded support, and (ii) whether the procedures are sufficiently robust with respect to contaminations. To allow comparisons with other procedures the critical value was chosen to ensure that the procedure achieves an in-control average run length of $\xi = 60$. Remaining parameters were optimized with respect to a unit shift of the mean. Optimizing for a unit shift (moderate jump) was considered to be a good compromise, when the aim is to evaluate the procedures for both small jumps, where EWMA charts are considered to be a good choice, and large jumps. The EWMA chart is given by $Z_n = (1 - \lambda)Z_{n-1} + \lambda Y_n$, and $\lambda = 1$ yields the Shewhart chart, which is preferable to detect large jumps. Optimization was done for λ between 0.01 and 0.99. For the (clipping) median the parameter M was also chosen to minimize the out-of-control ARL. We considered both the ARL and the probability that the procedure detects the change-point with no delay for various alternatives. Robustness was studied by a symmetric contaminated normal model. For each setting of the parameters of the simulation models 50,000 runs were used to estimate these quantities of interest.

Normal Errors: In our first experiment we generated series of i.i.d. $N(0,1)$–distributed error terms ϵ_n. In-control series were obtained by putting $Y_n = \epsilon_n$, whereas for out-of-control series an alternative $m(t)$ was added starting from time $n = 1$, i.e.,

$$Y_n = m(n) + \epsilon_n, \qquad n \in \mathbb{N}.$$

Given a generic pattern function m_0 defined on the unit interval $[0,1]$ the alternative m was defined as $m(n) = m_0(n/\xi)$ if $n \leq \xi$ and $m(n) = 0$ if $n > \xi$. This means, the pattern was stretched out onto the interval $[0, \xi]$.

For the clipping median we used an Epanechnikov kernel, and for the MedMin procedure we simply used the kernel

$$k(z) = \begin{cases} k_{\min} + 0.75 \cdot (1 - z^2), & |z| \leq 1 \\ k_{\min}, & |z| > 1. \end{cases}$$

obtained by adding the constant $k_{\min} = 0.5$.

Table 1 reports out-of-control ARLs and the first 4 atoms of the run length distribution for the ClipMed, the MedMin, and the EWMA procedures. We used $h = 5$ and $h = 10$. Which procedure is better depends on the alternative m_0. We studied two alternatives where the (clipping) was not expected to be perfect to avoid trivial results. As expected, in terms of ARL the EWMA is preferable. But it is worse in term of probability of no delay.

Noting that the first two alternatives start with a level shift of size 1 (moderate shift), for which the methods were optimized. The first one then decrease to 0, whereas the second increases. Although the EWMA is preferable in terms of ARL, it almost never detects the jump immediately, whereas the corresponding probability for the clipping median is 100 times higher. The third alternative, te^{-4t}, starts with a trend. In this situation the ClipMed and MedMin procedures are only slightly worse than the EWMA in terms of ARL.

The question arises, how the ARL and prob. of delay behave as a function of the jump height $a \geq 0$. Figure 1 provides the corresponding ARL curve for a pure jump model $m(n) = a$, whereas Figure 2 shows the corresponding prob. of no delay. If we are interested in detecting jumps immediately with high probability, the intuitive benefits of the clipping median is confirmed by our simulations.

Contaminated Normal Errors: Our second experiment concerns the robustness properties of the procedures examined in this article. In many applications it is reasonable to assume that a certain percentage, say, $\gamma \cdot 100\%$ of the data point are 'bad' ones. Typically, these gross errors inflate the variance and produce outliers which severely affect performance properties of classical procedures. To gain some insight into that question we used the same model as above but generated errors according to a symmetric contaminated normal distribution, i.e.,

$$\epsilon_n \sim \frac{\gamma}{2} N(-m_c, \sigma_c^2) + (1-\gamma) N(0, \sigma^2) + \frac{\gamma}{2} N(m_c, \sigma_c^2). \tag{12}$$

We used $\gamma = 0.1,\ m_c = 4$, and $\sigma_c = 1$. To provide a fair comparison, the threshold c was again chosen to ensure an in-control ARL 60, and remaining parameters were chosen to ensure detection of a one-unit level shift with smallest ARL. Table 2 provides the results. Whereas the run length distribution of the EWMA procedure is now severely shifted to the right, the respective distributions of the ClipMed and MedMin approaches are much less affected by the contamination. The loss of robustness of the EWMA procedure even when designed for a known contamination model as (12) results in a severe breakdown of the performance both in terms of ARL and probability of no delay function. Figure 3 shows that the benefits of the EWMA in terms of ARL now disappear. But Figure 4 shows that the clipping median is much more better in detecting the change with no delay. Whereas the curve of the EWMA procedure severely drops down, the curve of the clipping median is only slightly affected.

6. Conclusions

We studied a sequential detection rule based on a (clipping) median to ensure both robust smooths and immediate detection of jumps with high probability. The method is based on shrinking the observations towards 0 and a clipping mechanism neglecting observations far away from the current level. Sufficient conditions for zero delay are established for general clipping location estimators for arbitrary Banach space-valued data. For the case of no clipping we provide a sufficient condition for an asymptotic upper bound for the delay. The case of small (effective) sample sizes is studied by a simulation study. The clipping median seems to be preferable for many alternatives including the pure jump model, when interest focuses on robustness and the probability of no delay, whereas the performance of the EWMA severely breaks down for contaminated data. In particular, the benefit of the clipping median in terms of the probability of no delay drastically increases.

References

[1] Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics*, **37**, 577-580.

[2] Brodsky B. E. and Darkhovsky B. S. (1993), *Nonparametric Methods in Change-Point Problems*, Mathematics and its applications, **243**, Kluwer Academic Publishers, Dordrecht.

[3] Brodsky, B. E. and Darkhovsky B. S. (2000). *Non-Parametric Statistical Diagnosis Problems and Methods*. Mathematics and its applications, **509**, Kluwer Academic Publishers, Dordrecht.

[4] Chiu, C. K., Godtliebsen, F, and Marron, J. S. (1998). Edge-preserving smoothers for image processing. *Journal of the American Statistical Association*, **93**, 526-556.

[5] Girshick, M.A. and Rubin, H. (1952). A Bayes approach to a quality control model. *Annals of Mathematical Statistics*, **23**, 1, 114-125.

[6] Hesse, C. H. (1990). A Bahadur-type representation for empirical quantiles of a large class of stationary, possibly infinite-variance, linear processes. *Annals of Statistics*, **18**, 1188-1202.

[7] Kiefer, J. (1967). On Bahadur's representation of sample quantiles. *Annals of Mathematical Statistics*, **38**, 1323-1342.

[8] Kryzak A., Rafajłowicz E., Skubalska-Rafajłowicz E. (2001). Clipped median and space-filling curves in image filtering, *Nonlinear Analysis*, **47**, 303-314.

[9] Lee, J. S. (1983). Digital image smoothing and the sigma filter. *Computer Vision, Graphics, and Image Processing*. **24**, 255-269.

[10] Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, **14**, 1379-1387.

[11] Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, **1**, 100-115.

[12] Page, E.S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523-526.

[13] Pawlak, M. and Rafajłowicz, E. (2000). Vertically weighted regression - a tool for nonlinear data analysis and constructing control charts. *Journal of the German Statistical Association*, **84**, 367-388.

[14] Pawlak, M. and Rafajłowicz, E. (2001b). Jump preserving signal reconstruction using vertical weighting. *Nonlinear Analysis*, **47**, 327-338.

[15] Pawlak, M. and Rafajłowicz, E. (2001c). On a nonlinear method of function recovering: The vertical regression approach. Presented at: *Workshop on Nonlinear Estimation & Classification*, University of California, Berkeley.

[16] Pollak, M. (1978). Optimal detection of a change in distribution. *Annals of Statistics*, **13**, 206-227.

[17] Rafajłowicz, E. (1996). Model free control charts for continuous time production processes. Polytechnika Wrocław, *Technical Report* No. 56/96.

[18] Ritov, Y. (1990). Decision theoretic optimality of the CUSUM procedure. *Annals of Statistics*, **18**, 1464-1469.

[19] Rue H., Chu C.K., Godtliebsen F. and Marron J.S. (2002). M-smoother with local linear fit. *Journal of Nonparametric Statistics*, **14**, 155-168.

[20] Schmid, W. and Schöne, A. (1997). Some properties of the EWMA control chart in the presence of autocorrelation. *Annals of Statistics*, **25**, 3, 1277-1283.

[21] Schmid, W. and Steland, A. (2000). Sequential Control of Non-Stationary Processes by Nonparametric Kernel Control Charts. AStA, **84**, 3, 315-336.

[22] Serfling, R. J. (1980). *Approximation Theorems in Mathematical Statistics*. Wiley, New York.

[23] Shiryaev, A. N. (1963). On optimal methods in the quickiest detection problems. *Teor. Veroyatn., Primen.*, **8**, 26-51. (Russian)

[24] Siegmund, D. (1985). *Sequential Analysis*. Springer, New York.

[25] Skubalska-Rafajłowicz E. (1994). The closed curve filling multidimensional cube. *Technical Report*, **46/94**, ICT Technica University of Wrocław.

[26] Steland A. (2002a). Nonparametric monitoring of financial time series by jump-preserving estimators. *Statistical Papers*, **43**, 361-377.

[27] Steland A. (2002b). Sequential credit risk: Can we benefit from sequential nonparametric control? *Selected Papers Operation Research Proceedings 2002*, 337-344.

[28] Steland A. (2003a). Optimal sequential kernel smoothers for dependent processes. *Journal of Statistical Planning and Inference*, in revision.

[29] Steland A. (2003b). Sequential control of time series by functionals of kernel-weighted empirical processes under local alternatives. to appear in *Metrika*.

[30] Steland A. (2003c). NP-optimal kernels for nonparametric sequential detection rules. *Economic Quality Control*, submitted.

[31] Vasipoulos, A. V. and Stamboulis, A. P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Control*, **10**, 20-30.

[32] Yakir, B. (1996). A lower bound on the ARL to detection with a probability constraint on false alarm. *Annals of Statistics*, **24**, 431-435.

[33] Yakir, B. (1997). A note on optimal detection of a change in distribution. *Annal of Statistics*, **25**, 2117-2126.

Alternative	Method	ARL	Run Length Distribution			
			0	1	2	3
	MedClip 5	10.18314	0.124	0.105	0.090	0.079
	MedMin 5	8.51214	0.044	0.058	0.113	0.113
$exp(-t)$	MedClip 10	14.69552	0.117	0.093	0.079	0.066
	MedMin 10	20.089	0.036	0.033	0.033	0.031
	EWMA	2.33504	0.089	0.312	0.250	0.152
	MedClip 5	5.73346	0.134	0.114	0.101	0.089
	MedMin 5	5.94268	0.043	0.063	0.125	0.127
$1+t$	MedClip 10	7.27214	0.122	0.101	0.086	0.079
	MedMin 10	13.73018	0.035	0.031	0.031	0.032
	EWMA	2.09836	0.098	0.331	0.259	0.154
	MedClip 5	53.84614	0.021	0.020	0.021	0.020
	MedMin 5	52.12082	0.021	0.021	0.020	0.021
$texp(-4t)$	MedClip 10	55.73406	0.023	0.021	0.023	0.023
	MedMin 10	54.52496	0.023	0.022	0.022	0.021
	EWMA	39.37372	0.011	0.051	0.065	0.061

TABLE 1. Run length distribution for normal errors.

Alternative	Method	ARL	Run Length Distribution			
			0	1	2	3
	MedClip 5	10.354	0.116	0.105	0.093	0.078
	MedMin 5	9.745	0.044	0.058	0.083	0.103
$exp(-t)$	MedClip 10	10.354	0.116	0.105	0.093	0.078
	MedMin 10	11.693	0.033	0.035	0.046	0.046
	EWMA	8.008	0.001	0.078	0.131	0.127
	MedClip 5	5.797	0.124	0.121	0.102	0.092
	MedMin 5	6.782	0.043	0.061	0.092	0.112
$1+t$	MedClip 10	5.797	0.124	0.121	0.102	0.092
	MedMin 10	8.760	0.033	0.035	0.038	0.051
	EWMA	5.518	0.002	0.088	0.147	0.145
	MedClip 5	55.417	0.018	0.020	0.019	0.025
	MedMin 5	52.191	0.024	0.019	0.025	0.021
$texp(-4t)$	MedClip 10	55.417	0.018	0.020	0.019	0.025
	MedMin 10	52.437	0.027	0.027	0.023	0.023
	EWMA	51.392	0.000	0.006	0.018	0.021

TABLE 2. Run length distribution for contaminated normal errors.

Figure 1. *Normal errors: Out-of-control ARL as a function of the jump height. Shown are the EWMA chart (thin line), the MedClip 5 chart (bold line), and the MedMin 5 chart (dashed line).*

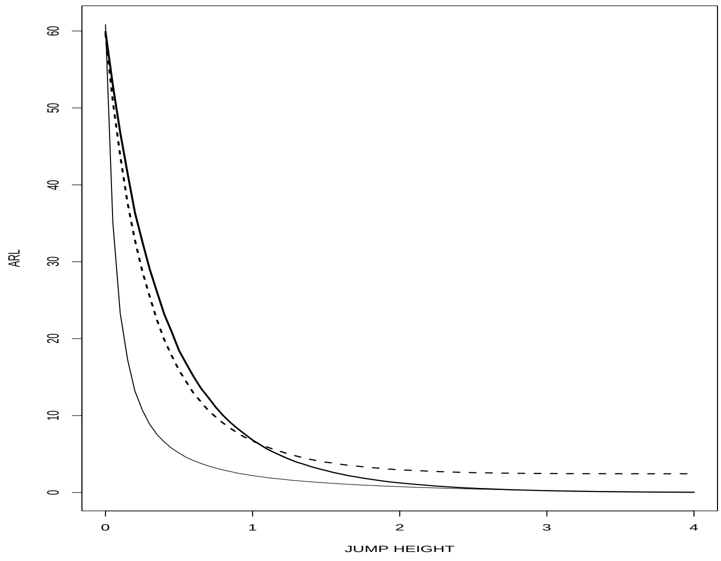

Figure 2. *Normal errors: The probability of no delay as a function of the jump height. Shown are the EWMA chart (thin line) and the MedClip 5 chart (bold line).*

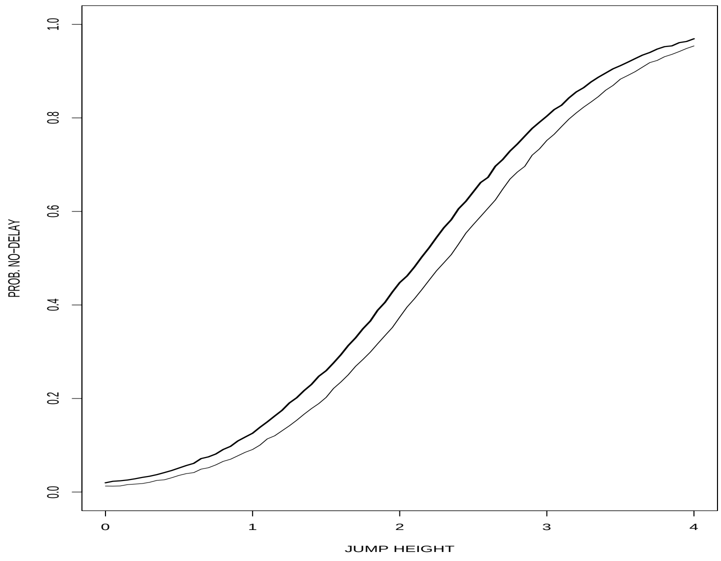

Figure 3. *Contaminated normal errors: Out-of-control ARL as a function of the jump height. Shown are the EWMA chart (thin line), the MedClip 5 chart (bold line), and the MedMin 5 chart (dashed).*

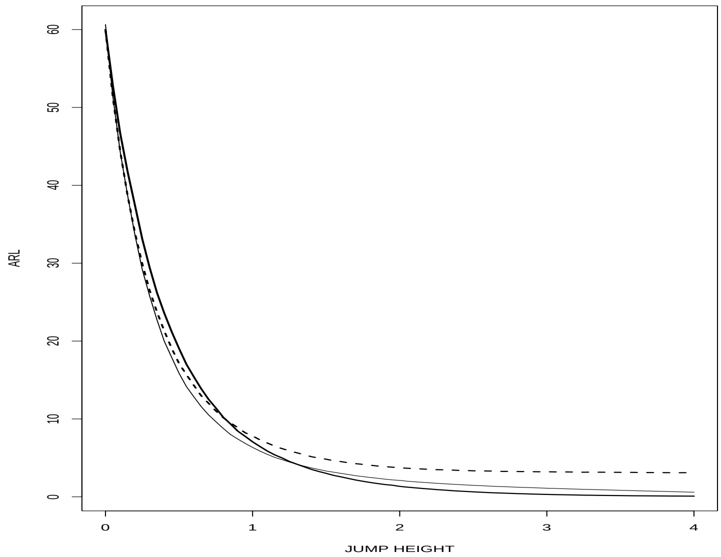

Figure 4. *Contaminated normal errors: The probability of no delay as a function of the jump height. Shown are the EWMA chart (thin line) and the MedClip 5 chart (bold line).*