



Pergamon

Nonlinear Analysis, Theory, Methods & Applications, Vol. 30, No. 7, pp. 3997–4006, 1997
Proc. 2nd World Congress of Nonlinear Analysts
© 1997 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0362-546X/97 \$17.00 + 0.00

PII: S0362-546X(97)00013-8

CHANGE-POINT MODELS IN INDUSTRIAL APPLICATIONS

EMMANUEL YASHCHIN

IBM Research Division, Mathematical Sciences Department, Thomas J. Watson Research Ctr.,
P.O. Box 218, Yorktown Heights, NY 10598, USA

Key words and phrases: Cusum, Detection, Filtering, Likelihood Ratio

1. INTRODUCTION

In many industrial applications of Statistics it is not reasonable to assume that the same model remains adequate as time progresses. Models in which the environment and related parameters undergo abrupt changes at unknown moments of time are found to be relevant in a much wider class of practical situations. These models spawned a number of fundamental problems in the field of the change-point theory, such as the problems of detection of changes (monitoring), estimation of the current process parameters (filtering), identifying points of change and regimes (segmentation) and tests for data homogeneity. These problems have been addressed, to various extent, in a large number of works, including several recent books and review papers (cf. [1–4]).

Problems related to change-point models are typically relevant in either fixed sample or sequential settings. For example, in the problem of on-line detection of a change decisions to trigger an out of control signal are made sequentially, based on some stopping variable. Some problems, however, can be formulated in both sequential and fixed sample settings. For example, in process capability analysis the problem of segmentation involves identifying all the regimes and change-points present in a given data set. However, in speech analysis the problem of segmentation is typically relevant in a sequential setting, with emphasis placed on identification of the most recent regime. Similarly, the problem of estimating parameters at a given point in time can be formulated as a sequential (filtering) or a fixed sample (smoothing) settings. In this article we focus on *sequential* methods, with emphasis on the problems of detection and filtering.

2. DETECTION

Let $\{\mathbf{X}_i\}$, $i = 1, 2, \dots$ be a sequence of (generally multivariate) observations that may represent, for example, proportions of defective items due to various causes, counts of contaminating particles of various types observed in successive periods of time, discrepancies between observed sales of a set of items and sales predicted by some model, and so forth. The stochastic behavior of the sequence is determined by the vector of parameters θ . For simplicity, we initially assume that all the components of this vector are of primary interest. Later in the text we address the case where behavior of $\{\mathbf{X}_i\}$ also depends on nuisance parameters. In general, the vectors $\{\mathbf{X}_i\}$ can form a serially correlated sequence; however, to simplify the presentation, in what follows we will assume independence except

where stated otherwise. Denote the most recent moment of time by T and the corresponding most recently observed observation by \mathbf{X}_T . Denote the joint distribution of m most recent observations by $f_{\theta}(\mathbf{X}_{T-m+1}, \dots, \mathbf{X}_{T-1}, \mathbf{X}_T)$, and denote its natural logarithm (the log-likelihood) by $L_m(\theta)$.

To set up the problem of detection, one should first specify the acceptable region Ω_0 in which θ should reside under normal operating conditions and the unacceptable region, Ω_1 . Note that the union of Ω_0 and Ω_1 does not need, in general, to cover the whole parameter space: there will generally exist a "grey" area in between. This three-zone approach is motivated by practical convenience: in many industrial applications an engineer will have no difficulty specifying areas that are distinctly "good" or "bad"; however, dividing the parameter space into two regions to separate "good" values from "bad" ones could prove to be a challenge.

Performance of detection schemes is typically measured in terms of the Run Length (RL), a random variable representing the number of observations taken until a signal is triggered. In general, one would like this variable to be large when $\theta \in \Omega_0$ (i.e., a low false alarm rate) and small when $\theta \in \Omega_1$ (i.e. good sensitivity with respect to out-of-control conditions). The most popular measure is the Average Run Length (ARL); however, design of control schemes based on quantiles and probabilities of RL has also been developed and supported in public domain software (see [5]).

Now let us denote, for a given set of last m observations,

$$L_{m0}^* = \max_{\theta \in \Omega_0} L_m(\theta), \quad L_{m1}^* = \max_{\theta \in \Omega_1} L_m(\theta), \quad D_m^* = L_{m1}^* - L_{m0}^*. \quad (2.1)$$

Denote by m_0 the minimal depth of data m for which θ is estimable. Then one can define a general strategy that leads to powerful control schemes as follows.

Likelihood Ratio (LR) Strategy: Trigger an out of control signal at time T if $D_m^* > h$ for some $m \geq m_0$ and pre-specified threshold h .

The above strategy leads to powerful procedures for a wide class of situations involving control of univariate and multivariate processes with or without serial correlation (see [2], [4]). Since it is not convenient for practical use (as it requires one to examine the whole data set to reach a decision whether a signal is to be triggered at time T), any practical application of the LR strategy involves choosing a window of size M and triggering a signal only if $D_m^* > h$ for values $m_0 \leq m \leq M$. In effect, this amounts to running a truncated SPRT backwards in time. As shown in [6], in the univariate case one can achieve asymptotic efficiency of the LR test by examining only a subset of values m . However, this approach could still require a search going deep into history to establish whether a signal is to be triggered.

In [7] an alternative procedure, the Regenerative Likelihood Ratio (RLR) is introduced. This procedure calls for determining the depth M_T dynamically, based on the previous history:

Regenerative Likelihood Ratio (RLR) Strategy: Given that at time T the last regeneration point was registered M_T units of time ago, trigger a signal if $D_m^* > h$ for some $m_0 \leq m \leq M_T$. If $D_m^* \leq 0$ for every $m_0 \leq m \leq M_T$, declare T the new regeneration point.

When the observations are univariate and their distribution belongs to an exponential family, the LR and RLR strategies are equivalent (see [7]). The latter paper also discusses the relative merits of these approaches. In what follows, we limit our attention to the RLR schemes.

In some applications one may want to consider simplified versions of the RLR strategy. One possibility is to examine only selected values of m in the interval $m_0 \leq m \leq M_T$ (for example, every k -th value) in the process of deciding whether a signal is to be triggered at time T . This point in time is declared the new regeneration point if $D_m^* \leq 0$ for every m on the k -spaced grid. When $k \rightarrow \infty$ this policy essentially amounts to running a sequence of SPRT tests and declaring a point at which $H_0 : \theta \in \Omega_0$ is accepted to be a new regeneration point. In general, it is advisable to include m_0

in the set of values of m to be examined - this will assure that changes of large magnitude will be detected quickly. Thus, one can consider using another simple scheme that examines only two values of m , namely, m_0 and M_T , and declares T the new regeneration point if $D_m^* \leq 0$ for both these values. One can see that in the case where both Ω_0 and Ω_1 contain only one value (the simple vs. simple hypothesis case) this approach scheme is very similar to the well known Cusum - Shewhart scheme (see [8], [5]), but differs from the latter in one important aspect that greatly simplifies the design process: it has only one parameter (h) instead of two. Other simplified forms of the RLR strategy can be found in [7].

In general, when considering a simplified RLR scheme, one should take in consideration not only sacrifices this will require in terms of the ARL curve (they frequently turn out to be quite tolerable), but also the way in which nuisance parameters are handled in the particular application. Let us denote the vector of nuisance parameters by η . Then we can define

$$L_{m_0}^* = \max_{\theta \in \Omega_0, \eta} L_m, \quad L_{m_1}^* = \max_{\theta \in \Omega_1, \eta} L_m, \quad D_m^* = L_{m_1}^* - L_{m_0}^*, \quad (2.2)$$

and apply an RLR scheme in the way described above (the value m_0 will have to be large enough to enable one to estimate not only θ but also η). In this mode we will assure that even an abrupt change in η will not prevent us from detecting unfavorable changes in θ reasonably fast; simplified RLR, however, could be much slower in detecting changes in θ under such circumstances.

In situations where changes in η tend to be infrequent, one can chose to obtain an estimate of its current value at time T and treat it as a known quantity. In the process of estimating the current value of η one will typically use (explicitly or implicitly) data that extend beyond the window M_T ; for example, one can use exponentially weighted averages or other filtering techniques. Several such techniques are discussed in the sections 4 - 6. In situations of this kind simplified RLR schemes tend to be less vulnerable.

The above summary only relates to the frequentist approach to the problem of detection. In the literature one can find a number of techniques that stem from the Bayesian approach to this problem that cannot be discussed here because of the limited scope of this article (e.g., see [3], [9] for information and references on this topic).

3. EXAMPLE: MONITORING A VARIANCE COMPONENT

Consider the problem of monitoring characteristics of oxide thickness in the process of manufacturing Integrated Circuits (chips) used in computing devices. Chips are typically processed as part of a wafer (thin disk about 20 cm in diameter; a wafer contains approximately 200 square shaped chips). Consider a specific process step which deposits a thin layer of silicon oxide onto the surface of a wafer. In this step, wafers are typically handled in lots of about 20. To monitor oxide thickness a sample of R wafers is randomly selected from each lot and N measurements are taken from each wafer. We assume that these measurements follow a nested random effect model,

$$X_{irn} = \mu + L_i + W_{r(i)} + E_{irn}, \quad i = 1, 2, \dots, \quad r = 1, 2, \dots, R, \quad n = 1, 2, \dots, N, \quad (3.1)$$

where X_{irn} is the thickness corresponding to the n -th site on the r -th wafer of the i -th lot, μ is the grand process mean, $L_i \sim N(0, \sigma_b)$ is the random effect of the i -th lot, $W_{r(i)} \sim N(0, \sigma_w)$ is the nested effect of the r -th wafer in the i -th lot and $E_{irn} \sim N(0, \sigma)$ is the random noise representing the effect of the n -th measurement taken from the r -th item of the i -th lot.

Let us focus our attention on monitoring the wafer-to-wafer component of variability, σ_w . In most applications, Ω_0 and Ω_1 are of the form $\sigma_w \leq \sigma_{w0}$ and $\sigma_w \geq \sigma_{w1}$, respectively, where $\sigma_{w0} < \sigma_{w1}$ are fixed based on engineering considerations. In this case the nuisance parameter is σ ; to simplify the

notation, we use it in a different form, namely, $\eta = \sigma^2/N$ which represents the part of the variance of the wafer averages that is explained by the within-wafer variability. Monitoring of σ_w is based on the sequence of bivariate statistics, $\{\hat{\sigma}_{i\bullet}^2, \hat{\sigma}_i^2\}$ defined by

$$\hat{\sigma}_{i\bullet}^2 = \frac{1}{R-1} \sum_{r=1}^R (\bar{X}_{ir\bullet} - \hat{\mu}_i)^2, \quad \hat{\sigma}_i^2 = \frac{1}{R(N-1)} \sum_{r=1}^R \sum_{n=1}^N (X_{irn} - \bar{X}_{ir\bullet})^2, \quad (3.2)$$

where $\hat{\mu}_i = \frac{1}{RN} \sum_{r=1}^R \sum_{n=1}^N X_{irn}$ and $\bar{X}_{ir\bullet}$ is the average of the N measurements taken from the r -th wafer of the i -th lot.

The log-likelihood based on the last m lots is given by

$$L_m(\sigma_w, \eta \mid \hat{\sigma}_{i\bullet}, \hat{\eta}_i, i = T-m+1, \dots, T) \propto C - mv_1 [\ln(\sigma_w^2 + \eta) + M_1/(\sigma_w^2 + \eta)] - mv_2 [\ln \eta + M_2/\eta] \quad (3.3)$$

where C does not depend on the parameters,

$$M_1 = \frac{1}{m} \sum_{i=T-m+1}^T \hat{\sigma}_{i\bullet}^2 \stackrel{\text{dist}}{=} (\sigma_w^2 + \eta) V_1[mv_1], \quad M_2 = \frac{1}{m} \sum_{i=T-m+1}^T \hat{\sigma}_i^2/N \stackrel{\text{dist}}{=} \eta V_2[mv_2].$$

are sufficient statistics related to σ_w and η , $v_1 = R-1$ and $v_2 = R(N-1)$ are degrees of freedom associated with $\hat{\sigma}_{i\bullet}^2$ and $\hat{\sigma}_i^2$, and $V_j[v]$ is a Chi-square random variable with v degrees of freedom divided by v (note that V_1 and V_2 are independent).

Now let us introduce the function $\eta^*(\sigma_w)$ which, for every σ_w returns the value η that maximizes the likelihood. Then one can show (see [7]) that L_{m0}^* and L_{m1}^* are determined as follows:

- (a) If $\hat{\sigma}_w \leq \sigma_{w0}$ set $L_{m0}^* = L_m(\hat{\sigma}_w, \hat{\eta})$ and $L_{m1}^* = L_m(\sigma_{w1}, \eta^*(\sigma_{w1}))$.
- (b) If $\sigma_{w0} < \hat{\sigma}_w \leq \sigma_{w1}$ set $L_{m0}^* = L_m(\sigma_{w0}, \eta^*(\sigma_{w0}))$ and $L_{m1}^* = L_m(\sigma_{w1}, \eta^*(\sigma_{w1}))$.
- (c) If $\hat{\sigma}_w > \sigma_{w1}$ set $L_{m1}^* = L_m(\hat{\sigma}_w, \hat{\eta})$ and $L_{m0}^* = L_m(\sigma_{w0}, \eta^*(\sigma_{w0}))$.

Now construction of an LR (or RLR) scheme is straightforward. The computations are greatly simplified by the fact that in the case (a) the score D_m^* is negative and thus it does not need to be computed. An example illustrating application of this scheme can be found in [7].

4. FILTERING IN THE PRESENCE OF ABRUPT CHANGES

In many practical situations involving abrupt changes in the process parameters the most important problem is not detection but rather estimation of the current level of parameters, i.e., filtering. This situation typically arises when one is able to neutralize the effect of unfavorable changes by making adjustments to the process. For example, consider the problem of contamination control in the process of chip manufacturing. Periodically, test wafers are introduced in various processing stages with the purpose of examining the contaminating particles that land on their surface. Detection schemes are then used to trigger a signal once particle counts become "large". However, process steps are typically followed by washing the wafers; the washing regime can be adjusted based on the current level of contamination. In this situation the problem of filtering is clearly of primary importance, even though the problem of detection also cannot be neglected.

Though the filtering problem can be formulated independently of the nature of θ , in this section we limit ourselves to the case where the parameter of interest is the mean of \mathbf{X}_i ; only the method for selecting the last stable regime from the sequence $\{\mathbf{X}_i\}$ will be formulated in a more general form because of its special importance. In the next section we give a brief summary of the case when θ is a vector of regression slopes.

The filtering problem for the mean has been discussed in a number of works (see [10] – [17]). Typically, this problem has been addressed within a Bayesian framework, with some prior distributions associated with location and magnitude of changes. In [16]-[17] a non-Bayesian approach is introduced. This approach is based on the loss function, for example,

$$L(\hat{\mu}_i, \mu_i) = \|\hat{\mu}_i - \mu_i\|_{\Sigma}^2 = (\hat{\mu}_i - \mu_i)^T \Sigma_i^{-1} (\hat{\mu}_i - \mu_i), \tag{4.1}$$

where $\hat{\mu}_i$ and μ_i are estimated and true means of \mathbf{X}_i , and Σ_i is its covariance matrix. The approach first limits the attention to the class of estimation schemes $\{\hat{\mu}_i\}$ that generate the same loss per unit time in the absence of abrupt changes. Within this class, best schemes are chosen based on the concept of *inertia* which measures the excess loss associated with estimation following an abrupt change that is due to absence of a-priori knowledge about its location and magnitude.

In what follows we assume that Σ_i is known and equal to Σ (generalizations to the case of unknown Σ_i can be found in [16]-[17]). In both cases the optimal linear scheme under the stated conditions is the multivariate Exponentially Weighted Moving Average (EWMA), see [17]. However, if changes in μ are not a-priori known to be of small or moderate value in relation to Σ , the EWMA schemes cannot be recommended for use: for example, they can produce estimates that are inconsistent with the most recent observations. This drawback can be eliminated by switching to Markovian schemes defined by a suitable initial value $\hat{\mu}_0$ and the process

$$\hat{\mu}_i = \mathbf{X}_i + \varphi(\|\hat{\mu}_{i-1} - \mathbf{X}_i\|_{\Sigma}) (\hat{\mu}_{i-1} - \mathbf{X}_i), \tag{4.2}$$

where $0 \leq \varphi(z) \leq 1$ is some non-increasing function defined for $z \geq 0$. Two functions recommended in [17] are defined in terms of two parameters:

$$\varphi_1(z) = \begin{cases} \gamma & z \leq c/\gamma \\ c/z & z > c/\gamma \end{cases} \tag{4.3a}$$

and

$$\varphi_2(z) = \gamma \exp[-0.5(z/\beta)^2]. \tag{4.3b}$$

The parameter c (or β) is used to determine the shape of the inertia function with respect to changes of various magnitudes and γ is a smoothing parameter that is used to establish the desired level of steady state loss per unit time in the absence of changes. In most practical applications γ is chosen between 0.7 and 0.95. Note that when $\varphi(z) \equiv z$ the scheme (4.2) reduces to the multivariate EWMA scheme.

When changes of large magnitude are common, better procedures are obtained by estimating, at any point T , the value of r_T , which represents the number of observations taken after the last point of change in the parameter of interest, θ (as noted earlier, in discussing this subject we consider a general parameter, not necessarily the mean; the formulation will also take into account the possibility of serial correlation in the sequence of observations). To be more precise, the estimate \hat{r}_T will specify the number of observations that are declared to be relevant when computing the current level of θ . The value \hat{r}_T will be called the *last stable range* (LSR) estimate.

The procedure for obtaining \hat{r}_T introduced in [16]-[17] proceeds sequentially from the current point T back into history and stops when the likelihood ratio test rejects the hypothesis that the last data segment can be explained by a single level of θ . At this point, all the preceding observations are ignored and \hat{r}_T is estimated.

Denote

$$M(n; T) = \max_{\theta_0, 1 \leq r \leq n-1, \theta_1} \log f(\mathbf{X}_T, \mathbf{X}_{T-1}, \dots, \mathbf{X}_{T-n+1} | \mathbf{X}_{T-n}, \dots, \mathbf{X}_1; \theta_0, r, \theta_1), \tag{4.4}$$

where f is the conditional density of the last n observations computed under the assumption that the last r observations correspond to the parameter θ_0 and all the preceding observations correspond to the parameter θ_1 . The value of r for which the maximum is achieved is denoted by $r(n; T)$. Similarly, denote

$$M_0(n; T) = \max_{\theta} \log f(\mathbf{X}_T, \mathbf{X}_{T-1}, \dots, \mathbf{X}_{T-n+1} | \mathbf{X}_{T-n}, \dots, \mathbf{X}_1; \theta), \quad (4.5)$$

where f is the conditional density of the last n observations computed under the assumption that all of them correspond to the parameter θ . Now the procedure for deriving \hat{r}_T is formulated as follows:

Procedure A: Select a positive threshold h and, for $n = 2, 3, \dots$, compute

$$d(n; T) = M(n; T) - M_0(n; T), \quad (4.6)$$

until for the first time $d(n; T) > h$. At this point, set $\hat{r}_T = r(n; T)$. If $n = T$ is reached and $d(T; T) \leq h$, set $\hat{r}_T = T$.

Of special interest is the case where $\{\mathbf{X}_i\}$ are independent Gaussian with mean μ_i and known Σ , and the filtered parameter is the process mean. In this case the procedure of obtaining \hat{r}_T is especially simple since

$$d(n; T) = \max_{1 \leq r \leq n-1} \frac{r(n-r)}{2n} \|\bar{\mathbf{X}}_T(r) - \bar{\mathbf{X}}_{T-r}(n-r)\|_{\Sigma}^2, \quad (4.7)$$

where $\bar{\mathbf{X}}_T(r) = r^{-1} \sum_{j=1}^r \mathbf{X}_{T-j+1}$.

Once \hat{r}_T is obtained, one can use the following procedure to obtain $\hat{\mu}_T$, which is defined in terms of two parameters, the threshold $h > 0$ and smoothing parameter, $0 < \gamma \leq 1$:

Adaptive Exponentially Weighted (AEW) Scheme:

Step1: Find \hat{r}_T by using Procedure A;

Step2: If $\hat{r}_T = \hat{r}_{T-1} + 1$ then set

$$\hat{\mu}_T = \hat{\mu}_{T-1} + \frac{1-\gamma}{1-\gamma^{\hat{r}_T}} (\mathbf{X}_T - \hat{\mu}_{T-1}). \quad (4.8)$$

Otherwise, set

$$\hat{\mu}_T = \frac{(\mathbf{X}_T + \gamma \mathbf{X}_{T-1} + \dots + \gamma^{\hat{r}_T} \mathbf{X}_{T-\hat{r}_T+1})}{(1 + \gamma + \dots + \gamma^{\hat{r}_T-1})} \quad (4.9)$$

Note that all the schemes discussed in this section reduce to the multivariate EWMA with parameter γ when the second parameter (c , β or h) tends to infinity.

Several examples illustrating use of the above methods in problems related to chip fabrication can be found in [16]–[17].

5. WEIGHTED LIKELIHOOD AND ITS APPLICATIONS

In this section we discuss generalizations related to weighting the observations. We introduce the weighted likelihood technique and show how it can be used in problems of detection and filtering. As one will see, by using this concept one can easily extend the methods used above to obtain filtered estimates of the process mean to cover the case of filtering a more general parameter, θ .

Let w_0, w_1, \dots, w_{m-1} be weights associated with $\mathbf{X}_T, \mathbf{X}_{T-1}, \dots, \mathbf{X}_{T-m+1}$, respectively, for any fixed m (in other words, w_0 is associated with the last observation, w_1 - with the previous one, etc.). The weights generally decrease to provide emphasis on the most recent information. Then the weighted log-likelihood function corresponding to the last m observations is given by:

$$L_m^{(w)}(\theta) = \sum_{i=T-m+1}^T w_{T-i} \log f_{\theta}(\mathbf{X}_i | \mathbf{X}_{i-1}, \mathbf{X}_{i-2}, \dots). \quad (5.1)$$

First, let us illustrate the use of weighted likelihood in the problem of detection. In light of the LR approach described in Section 2, one can construct weighted Likelihood Ratio control schemes by defining

$$L_{m_0}^{*(w)} = \max_{\theta \in \Omega_0} L_m^{(w)}(\theta), \quad L_{m_1}^{*(w)} = \max_{\theta \in \Omega_1} L_m^{(w)}(\theta), \quad D_m^{*(w)} = L_{m_1}^{*(w)} - L_{m_0}^{*(w)} \quad (5.2)$$

and triggering an out of control signal at time T if $D_m^{*(w)} > h$ for some $m \geq m_0$ and threshold h . Schemes of this type are called *weighted LR control schemes of type 2*. As shown in [3], schemes of this type enable one to improve performance with respect to *drifts* in θ with minimal performance loss with respect to *shifts*. In many practical applications it is convenient to choose weights of type $w_i = \gamma^i$, $i = 0, 1, \dots$ which lead to *Geometric LR schemes*. Weighted RLR schemes can also be easily constructed based on (5.2).

Now let us return to the problem of filtering. As we saw earlier, in the case of filtering μ a simple approach to this problem is to begin with the EWMA and modify it so as to obtain acceptable performance, in terms of inertia, with respect to larger changes. When the parameter of interest θ is not the process mean, the approach to filtering can be developed along similar lines. First of all, we begin with an estimating scheme that performs well when changes in θ are small or moderate, and then we modify this scheme to accommodate the possibility of larger changes.

To obtain an equivalent of the Weighted Moving Average scheme we introduce the concept of Weighted Maximum Likelihood (WML) estimation. In particular, the WML estimator $\hat{\theta}_T$ based on the last m observations is the value of θ that maximizes $L_m^{(w)}(\theta)$. When $m = T$ and weights are decreasing geometrically, the WML estimator can be viewed as an analog of the EWMA filter, and (with proper selection of weights) it leads to schemes that have a good performance, in terms of inertia, when changes in θ are small or moderate. Such an estimator will be called an EWML estimator.

Modification of the WML so as to obtain an estimator of the current value of θ that is capable of adapting to large changes in θ can now be achieved by using the following Principle of Consistency: *the density of the most recent observation X_T computed under the assumption that $\theta = \hat{\theta}_T$ must always be consistent with X_T in the sense that this observation does not fall too far into the tail of its estimated density*. In the case of estimating the current level of μ the above principle was enforced by supplementing the smoothing parameter γ with an additional parameter (c, β or h , depending on the chosen scheme). In the more general case of estimating θ one can also introduce such an additional scalar parameter. For example, an AEWML scheme (analogous to the AEW scheme of Section 4) can be easily constructed by combining the Procedure A with the WML estimate based on \hat{r}_T .

WML is not the only filtering technique which enables one to put a larger emphasis on the most recent information. In many cases one can also obtain Weighted Least Square (WLS) estimates by minimizing the sum of squares of weighted residuals. When weights associated with the residuals decrease geometrically, we call this method the Exponentially Weighted Least Squares (EWLS) estimation. The WLS estimators can be adjusted to accommodate large changes in θ by using the methods described in the previous paragraph. For example, in combination with the estimate \hat{r}_T of the last stable range, the above scheme will be called the AEWLS scheme.

Of special interest is the problem of estimating the current value of multivariate regression slopes considered in the next section.

6. ESTIMATING THE CURRENT REGRESSION SLOPES

Consider the multiple regression case, where the observations $\{y_i\}$ are coming from the model

$$y_i = \mathbf{x}_i' \beta_i + \epsilon_i, \quad i = 1, 2, \dots, \quad (6.1)$$

where \mathbf{x}_i is a vector representing the independent variables, β_i is the vector of slopes and ϵ_i is the noise (all the bold-letter vectors discussed below are k -dimensional column-vectors). For simplicity, we only consider here the case where measurements at different points in time are independent and ϵ_i is Gaussian with mean zero and variance that is known and equal to σ^2 for all i .

Denote the WML estimate of β_i at time T based on the last m observations by $\hat{\beta}_T(m)$. This estimate is obtained by minimizing the weighted error sum of squares,

$$L_m^{(w)}(\beta) = \sum_{i=T-m+1}^T w_{T-i} (y_i - \mathbf{x}_i' \beta_i)^2, \quad (6.2)$$

with respect to β (as one can see, in the Gaussian case the WML and WLS methods produce identical estimates). This results in a well known WLS estimate

$$\hat{\beta}_T(m) = (X'WX)^{-1} X'W(y_T, y_{T-1}, \dots, y_{T-m+1})', \quad (6.2)$$

where rows of X contain the vectors $\mathbf{x}_T', \mathbf{x}_{T-1}', \dots, \mathbf{x}_{T-m+1}'$, and W is the matrix of weights; in our simple case $W = \text{diag}(w_0, w_1, \dots, w_{m-1})$.

Of special interest is the case $w_i = \gamma^i$, $i = 0, 1, \dots$. In this case one can show that $\hat{\beta}_T(m)$ can be computed recursively, based on $\hat{\beta}_{T-1}(m-1)$ and the most recent observation. Denote $\mathbf{G}_T(m) = (X'WX)^{-1}$. This matrix is always stored in the process of recursion; thus, at time T we have $\mathbf{G}_{T-1}(m-1)$ and $\hat{\beta}_{T-1}(m-1)$ available. The updating process is then as follows:

EWLS filter for β_i :

Step 1: Compute the direction of change, $\mathbf{z}_T = \mathbf{G}_{T-1}(m-1) \bullet \mathbf{x}_T$. Define $\bar{\mathbf{z}}_T = \mathbf{z}_T(\gamma + \mathbf{x}_T' \mathbf{z}_T)^{-1}$.

Step 2: Update the estimate: $\hat{\beta}_T(m) = \hat{\beta}_{T-1}(m-1) + \bar{\mathbf{z}}_T \bullet [y_T - \mathbf{x}_T' \hat{\beta}_{T-1}(m-1)]$ (6.3)

Step 3: Update the inverse: $\mathbf{G}_T(m) = \mathbf{G}_{T-1}(m-1) - \bar{\mathbf{z}}_T \mathbf{z}_T'$.

This process holds also when the previous history is infinitely deep, i.e., $m = \infty$.

Unfortunately, the above process suffers from the same drawback as the EWMA: it tracks large changes slowly. To implement the Principle of Consistency stated in the previous section we can use several approaches; the simplest one is to adjust w_0 upwards until the predicted value of y_T falls within some tolerable distance (say, $c\sigma$) from the observed y_T . A choice of $c \approx 2$ is reasonable in many practical situations. In EWLS schemes one can apply the procedure as described above if

$$|y_T - \mathbf{x}_T' \hat{\beta}_T(m)| \equiv \gamma(\gamma + \mathbf{x}_T' \mathbf{z}_T)^{-1} |y_T - \mathbf{x}_T' \hat{\beta}_{T-1}(m-1)| \leq c\sigma. \quad (6.4)$$

Otherwise, we should include one additional step (following Step 2) in the above procedure:

Step 2a: If (6.4) is not satisfied, compute $\gamma^* = \mathbf{x}_T' \mathbf{z}_T \left[|y_T - \mathbf{x}_T' \hat{\beta}_T(m)| / (c\sigma) - 1 \right]^{-1}$ (6.5) and repeat Steps 1 to 3 with $\gamma = \gamma^*$. In subsequent estimation, however, use the original value of γ .

It is also not difficult to obtain an AEW version of the filtering scheme for β_i . In this scheme we always store the last stable range, \hat{r}_{T-1} , and the associated values $\mathbf{G}_{T-1}(\hat{r}_{T-1}-1)$ and $\hat{\beta}_{T-1}(\hat{r}_{T-1}-1)$.

AEWLS filter for β_i :

Step1: Find \hat{r}_T by using Procedure A;

Step2: If $\hat{r}_T = \hat{r}_{T-1} + 1$ then obtain $\hat{\beta}_T(\hat{r}_T)$ by executing Steps 1 - 3 of the procedure (6.3). Otherwise compute $\hat{\beta}_T(\hat{r}_T)$ by using (6.2); save $\mathbf{G}_T(\hat{r}_T)$ for future use.

It is not difficult to see that in the case of intercept - only model with $k = 1$ this procedure reduces to the AEW procedure for estimating the current mean given in Section 4.

Numerous issues arise in relation to use of the above techniques in practice. For example, one can show that the procedures for estimating μ_T when the variance is unknown can be generalized for the regression case. Another interesting issue worth mentioning is how to handle situations where $\hat{r}_T < m_0$, i.e., the last stable range is identified more or less correctly, but it does not contain enough data to estimate β_T . In situations like that there will always be exposure in terms of loss and a good strategy to follow depends on the nature of changes in β expected in the application of interest. In some cases one can even get away with refusing to produce an estimate until information sufficient for identification of the new regime becomes available. One can also consider using a Bayesian approach to handle such situations. Because of the limited scope of this article discussion on these issues will be omitted.

7. CONCLUSIONS

Situations in which data can be viewed as being generated by models with change - points are very common in industry, especially in areas related to Quality Control. In this article we presented several methods for handling two of the most important problems: detection of changes in θ and estimation of its current level. These methods are based on the concept of Likelihood and Likelihood Ratio and they do not require assumptions about the process of changes. Especially useful in practical applications are the RLR detection schemes that are statistically powerful, easy to design and to implement. In my opinion, this method is likely to play a major role as the industry moves from the statistical methods prevalent in the first half of this century to more modern techniques. It can be viewed as the natural generalization of the conventional Cusum schemes to cover more complex data models, such as multivariate time series.

To address the problem of filtering in the context of change - point models we propose several methods all of which can be viewed as the generalization of the basic EWMA technique. Our approach is similar, in spirit, to the Neyman - Pearson approach to hypothesis testing: first, we restrict our attention to the class of schemes that can be viewed as "equivalent" under steady state conditions, and then we select schemes that show the best tracking capability. The concept of LSR is very useful and I believe that the Procedure A formulated in Section 4 will catch on in the engineering community. An interesting question remains whether exponentially decreasing weights are the best ones to use in conjunction with the Procedure A, as suggested in the AEW scheme. This choice is definitely the easiest to implement because it typically results in simple recursive schemes and, therefore, one can expect it to be used extensively, even in some cases where it can be proven sub-optimal.

A large number of issues arise in relation to any given practical case where use of such techniques is considered. How to determine "good" and "bad" process windows? How to handle the nuisance parameters? Should any transformations be applied to the data? What are the relevant sources of variability? Is there any serial correlation present and, if yes, what is its origin and nature? What modifications are needed if presence of outliers cannot be ruled out? How to obtain good performance estimates? What actions to take when we are quite confident that we are into a new regime but there is not enough data to estimate its characteristics? Under what conditions should we consider Bayesian methods more suitable? In any more or less complex situations designing a robust monitoring system involves not only solid science but also a great deal of art.

REFERENCES

1. CARLSTEIN, E., MULLER, H. & SIEGMUND, D. (eds.), Change-Point Problems, *Lecture Notes - Monograph Series* **23**, Institute of Mathematical Statistics (1995).

2. BASSEVILLE, M. & NIKIFOROV, I., *Detection of Abrupt Changes - Theory and Application*, Prentice Hall, Englewood Cliffs, New Jersey (1994).
3. YASHCHIN, E., *Statistical Control Schemes: Methods, Applications and Generalizations*, *International Statistical Review* **61**, 41–66 (1994).
4. TELKSNYS, L. (ed.), *Detection of Changes in Random Processes*, Optimization Software, Inc., New York (1986).
5. YASHCHIN, E., On the analysis and design of Cusum-Shewhart control schemes, *IBM Journ. Res. Devel.* **29**, 377–391 (1985).
6. LAI, T. L., Sequential Changepoint Detection in Quality Control and Dynamic Systems, *J. Royal Stat. Soc. B* **57**, 613–658 (1995).
7. YASHCHIN, E., Likelihood Ratio Methods for Monitoring Parameters of a Nested Random Effect Model, *J. Amer. Statist. Assoc.* **90**, 729–738 (1995).
8. LUCAS, J., Combined Shewhart-Cusum Quality Control Schemes, *J. Qual. Technol.* **14**, 51–59 (1982).
9. BERK, R. H., RINOTT, J. & YAKIR, B. Efficient Detection of a Change-Point, *Proc. 2-nd World Congr. of IFNA*, (1996), to appear.
10. BARNARD, G., Control Charts and Stochastic Processes *J. Royal Stat. Soc. B* **21**, 239–257 (1959).
11. CHEN, C. & TIAO, G. C., Random Level-Shift Time Series Models, ARIMA Approximations and Level-Shift Detection, *J. Business and Econ. Stat.* **8**, 83–97 (1990)
12. CHERNOFF, H. & ZACKS, S. Estimating the Current Mean of a Normal Population which is Subject to Changes in Time, *Ann. Math. Stat.* **35**, 999–1018 (1964).
13. KENETT, R. & ZACKS, S., Tracking Algorithms for Processes with Change Points, *submitted for publication* (1992)
14. McCULLOCH, R. E. & TSAY, R. S., Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series, *J. Amer. Stat. Assoc.* **88**, 968–978 (1993).
15. WEST, M. Bayesian Model Monitoring, *J. Royal Stat. Soc. B* **48**, 70–78 (1986).
16. YASHCHIN, E., Estimating the Current Mean of a Process Subject to Abrupt Changes, *Technometrics* **37**, 311–323 (1995).
17. YASHCHIN, E., Estimating the Current Mean of Multivariate Processes Subject to Abrupt Changes, *Proc. Sect. Phys. & Eng. Sciences, Amer. Stat. Assoc.*, 81–86 (1994).