MST '94

Proceedings of
the International Symposium on

MANUFACTURING SCIENCE AND TECHNOLOGY FOR THE 21ST CENTURY

The First S.M.Wu Memorial Symposium

July 3—5, 1994
Tsinghua University, Beijing, China

Edited by
Zhou Zhaoying

INTERNATIONAL ACADEMIC PUBLISHERS

A STATISTICAL PROCESS CONTROL METHOD FOR AUTOCORRELATED DATA USING A GLRT

Daniel W. Apley and Jianjun Shi
S. M. Wu Manufacturing Research Laboratory
Dept. of Mechanical Engineering and Applied Mechanics
The University of Michigan
Ann Arbor, Michigan 48109-2125

ABSTRACT

This paper presents an on-line statistical testing procedure, based on a generalized likelihood ratio test (GLRT), for detecting and estimating faults in correlated processes. The process is assumed to be ARMA, and the faults are assumed to be additive (e.g. a step change in the mean of the process or a spike in the data). In addition to detecting the fault, the GLRT developed here estimates the fault magnitude and the time of occurrence of the fault and classifies the fault according to a prespecified set of fault types. In this sense the GLRT combines the tasks of fault detection, estimation, and classification. It is shown that the estimate of the fault magnitude is both unbiased and efficient. Simulation results, which demonstrate the effectiveness of the GLRT in detecting both step mean shifts and spikes, are presented.

1 INTRODUCTION

With increasing automation in manufacturing processes the case of 100% inspection, or near 100% inspection, is becoming more and more common. However, as the process is sampled at higher rates, the measurement data, to be used for quality control and/or process control purposes, is more likely to be autocorrelated. It is well known that the performance of conventional quality control techniques, e.g. cusum tests, deteriorates when applied to correlated data. More specifically, the probability of false alarm may be significantly increased. Consequently, when the data exhibits correlation some alternative form of statistical testing should be used, or, at the very least, the control limits of the conventional techniques should be modified.

Previously, a good deal of research has been directed towards the modification of conventional fault detection tests and the design of new tests for correlated data. The average run length (ARL) for cusum tests on correlated data has been investigated (Johnson and Bagshaw, 1974, and Yashchin, 1993). Vasilopoulos and Stamboulis (1978) develop modified control limits taking into account the correlation of the data. Alt et al. (1977) designs a maximum likelihood estimate of the mean vector for a multivariate process and derives appropriate control limits for detecting a mean shift.

An alternative approach is to whiten the correlated data and use conventional control charts on the uncorrelated residuals. Most approaches of this type assume the data can be described by an invertible ARIMA time series model driven by white Gaussian noise (see, for example, Box and Jenkins (1976) or Pandit and Wu (1990)). If this assumption holds then the whitening filter is simply the inverse time series model, and the residuals are the one-step-ahead prediction errors of the model. Since the residuals are approximately uncorrelated if the model is adequate, conventional control charts can then be applied to the residuals. Under this approach, the statistical test applied to the residuals generally falls under one of three types: (1) a cusum test for detecting changes in the process mean; (2) a chi-squared test for detecting changes in the variance of the driving noise of the system (assuming it is white and Gaussian); and (3) a whiteness test for detecting changes in the process parameters (i.e. natural frequencies and damping ratios) of the system. In Dooley et al. (1986) the residuals are tested using both a whiteness test and a cusum test. A cusum and chi-squared test have been used to detect changes in principal stress data on a blast furnace shell (Notohardjono and Ermer, 1986). A methodology incorporating all three above mentioned tests, in conjunction with rule-based classification, was developed and applied to force signals in an end milling process in Dooley and Kapoor (1990). In addition, the effect of the three types of faults considered (mean shift, parameter shift, and change in noise variance) on each of the three tests are discussed in Dooley and Kapoor (1990). Processes that can be approximated by EWMA models are considered in Montgomery and Mastrangelo (1991), where control limits for run charts are developed based on appropriate choices of the weighting factor, λ, in the EWMA model. It has been suggested (Montgomery and Mastrangelo, 1991) that run charts on the residuals should be accompanied by run charts on the original correlated data.

The statistical test for fault detection proposed in this paper is related to the above approaches in that it analyzes the uncorrelated residuals of the appropriate whitening filter. However, instead of using a combination of the three more conventional tests described in the preceeding paragraph, the new method is based on a generalized likelihood ratio test (GLRT). The GLRT method is an inherently attractive approach to fault detection because its theoretical optimality, albeit in an off-line setting, has been established in Deshayes and Picard (1986). Furthermore, for detecting mean shifts in the process, the GLRT has an advantage over a CUSUM applied to the residuals, since the CUSUM test ignores the dynamics of the ARMA model and assumes a step-change in the process results in a step change in the residuals. The GLRT is described in a very general context in Van Trees (1968). The concept of using a GLRT to detect system faults gained

attention in the mid seventies (Willsky and Jones, 1976). For a comprehensive survey on the early work see Willsky (1976). Basseville (1988) provides a more recent survey on fault detection in dynamic systems, in general. Most of the work on fault detection using GLRTs has been more in the context of system monitoring than quality control. Hence, the techniques have been applied to the detection of faults in sensors, actuators, and the system dynamics. The majority of these approaches use a state-space representation of the process, as opposed to an ARMA representation, and a Kalman Filter for generating the residuals. Tsay (1988) proposed what amounts to a GLRT for detecting mean shifts, spikes, and variance changes in time series. However, the algorithm is a batch method designed for the off-line analysis of data and is not suitable for implementation in an on-line SPC framework, unlike the method developed in this paper.

The purpose of this paper is to present a sequential GLRT based SPC method for correlated processes. The method developed in this paper combines the tasks of fault detection, classification, and estimation. Here, classification is in regard to fault type (e.g. a mean shift or spike), and estimation is in regard to time of occurrance and fault magnitude. Furthermore, the implementation is straightforward and computationally inexpensive.

The format of the paper is as follows. Section 2 provides a description of the assumed process and fault models. In sections 3.1 through 3.3 the GLRT algorithm is developed and analyzed, and in section 3.4 threshold and window length selection is discussed. In section 4, implementation issues are discussed, and simulation results for detecting both a mean shift and a spike are provided.

2 SYSTEM DESCRIPTION

Consider a process which can be described by the following linear, time-invariant, discrete-time ARMA(p,q) model:

$$x(t) = G(B)a(t) := \frac{\Theta(B)}{\Phi(B)}\,a(t) := \frac{1+\theta_1 B+\theta_2 B^2+\ldots+\theta_q B^q}{1+\phi_1 B+\phi_2 B^2+\ldots+\phi_p B^p}\,a(t), \quad (1)$$

where a(t) are identically distributed, zero-mean, white Gaussian noise with variance σ_a^2, and B is the backshift operator. In this paper we assume that the ARMA model, G(B), is known either through *a priori* information or using any of a number of parameter estimation methods performed during a no fault period of the process. In addition, it is assumed that the ARMA process is both asymptotically stable and invertible.

In the development of the GLRT it is assumed that the output of the correlated process to be monitored can be described by the following model:

$$y(t) = x(t) + Kf_{j,\tau}(t): \quad j = 1, 2, \ldots, m, \quad (2)$$

where $f_{j,\tau}(t)$ is a unit magnitude fault of the j^{th} type, occurring at time τ, K is the magnitude of the fault, and x(t) is the ARMA process of equation (1). In the model described by equation (2) let there be m different fault types hypothesized. As an example, suppose m = 2, where a step mean shift (j = 1) and a spike (j = 2) are the hypothesized faults. Then, in this case

$$f_{1,\tau}(t) := \begin{cases} 0: & t < \tau \\ 1: & t \geq \tau \end{cases}, \quad \text{and} \quad f_{2,\tau}(t) := \begin{cases} 0: & t \neq \tau \\ 1: & t = \tau \end{cases}.$$

A final assumption is that no more than one type of fault may occur at any given time, and that the occurrence of different faults are spaced far enough apart that their effects do not overlap.

Since the invertibility of G(B) is assumed, we can define the inverse transfer function, or whitening filter, as

$$G^{-1}(B) := \frac{\Phi(B)}{\Theta(B)}. \quad (3)$$

With $G^{-1}(B)$ a linear filter, if the process output y(t) is filtered by $G^{-1}(B)$ and the residuals denoted e(t), the result is, using equations (1) through (3),

$$e(t) = G^{-1}(B)y(t) = a(t) + K\tau_{j,\tau}(t), \quad (4)$$

where $\tau_{j,\tau}(t) := G^{-1}(B)f_{j,\tau}(t)$ is the response of a unit magnitude fault when filtered by $G^{-1}(B)$. $\tau_{j,\tau}(t)$ will therefore be referred to as the fault signature of $f_{j,\tau}(t)$, an explanation of which will be given in the following paragraphs. Note that under no fault ($t < \tau$) conditions e(t) = a(t) and is the error, or residual, of the linear minimum mean square error estimator of y(t) given y(t-1), y(t-2), y(t-3), . . .

Equation (4) is conceptually important in that it shows e(t) is the sum of the NID(0, σ_a^2) sequence a(t) and the deterministic fault signature. e(t) is thus an uncorrelated Gaussian sequence with mean $K\tau_{j,\tau}(t)$ and variance $\sigma_e^2 = \sigma_a^2$, where the variance is independent of the occurrence of a fault.

Note that if a step mean shift occurs in the process, the resulting mean shift in the residuals will not be a step function. It will be a filtered step response with dynamics dependent on G(B). If a conventional cusum test were used on the residuals, the information contained in the fault signature dynamics would be ignored. It is reasonable to assume that by making use of the fault signature dynamics, a better fault testing procedure can be designed. As will be shown in subsequent sections, the GLRT takes into consideration these dynamics and, as a result, is capable of improved detection performance.

3 THE GLRT ALGORITHM

3.1 A Likelihood Ratio Test (K Known)

In this section a likelihood ratio test (LRT), for which it is assumed K of equation (2) is known, is developed. In the subsequent sections the LRT will be extended to a GLRT, which applies to the more general situation where K must be estimated. We first formalize the problem by defining the set of statistical hypotheses to be tested. For implementation purposes, instead of testing for the occurrance of faults at all previous times, only faults occurring in the interval {t-N, t-N+1, . . ., t} will be tested for, where t is the current time and N+1 is the window length. In general, when selecting N there is a tradeoff between computational complexity and probability of detection. Guidelines for selecting N will be discussed in section 3.4.

The convention behind the hypotheses definitions is as follows. Suppose that m types of faults are hypothesized, that the window length is set at N+1, and let M := m(N+1). Then, at each time t, the following hypotheses would be tested:

the null hypothesis

$H_0(t)$: no fault has occurred,

and the alternative hypotheses

N+1 hypotheses associated with fault type 1:
- $H_1(t)$: a type 1 fault occurred at time t
- $H_2(t)$: a type 1 fault occurred at time t-1
- •
- •
- •
- $H_{N+1}(t)$: a type 1 fault occurred at time t-N

N+1 hypotheses associated with fault type 2:
- $H_{N+2}(t)$: a type 2 fault occurred at time t
- $H_{N+3}(t)$: a type 2 fault occurred at time t-1
- •
- •
- •
- $H_{2(N+1)}(t)$: a type 2 fault occurred at time t-N

•
•
•

N+1 hypotheses associated with fault type m:

- $H_{m(N+1)-N}(t)$: a type m fault occurred at time t
- $H_{m(N+1)-N+1}(t)$: a type m fault occurred at time t-1
- •
- •
- •
- $H_{m(N+1)}(t)$: a type m fault occurred at time t-N

The fault detection and estimation task then becomes determining which of the above M+1 hypotheses is most likely. To simplify the analysis, we now introduce the following notation.

Definition: (5)

At a given time t define the fault function $f_i^t(\cdot)$ as that occurring under $H_i(t)$ in the above hypotheses definitions and $\underline{f}_i^t(\cdot)$ as its corresponding fault signature (see equation (4)) when filtered by $G^{-1}(B)$. Also define K_i as the magnitude of the fault occurring under $H_i(t)$.

As mentioned earlier in this section, we are looking only for faults occurring in the interval {t-N, t-N+1, . . ., t}. Since, if τ is the time of occurrence of the fault, e(t) ($t < \tau$) are distrubuted independently of which hypothesis is true, for the detection problem we need only consider the N+1 length residual vector defined as

$$\underline{e}(t) := [e(t\text{-}N)\ e(t\text{-}N+1) \ldots e(t)]^T \tag{6}$$

Determining which of the M+1 hypotheses is most likely is a multiple hypotheses testing problem. If the fault magnitudes are known *a priori* then an LRT testing procedure, which is known to be optimal in the Neyman-Pearson sense (Van Trees, 1968), could be used. Assuming for the remainder of this section that the fault magnitudes are known, the LRT is developed as follows. At each time t, and for i = 1, 2, . . ., M, the likelihood ratios

$$\Lambda_i := \frac{p_{\underline{e}|H_i}(\underline{e}|H_i)}{p_{\underline{e}|H_0}(\underline{e}|H_0)} \tag{7}$$

are calculated. In equation (7) the numerator and denominator are the conditional probability densities of the given residual vector under H_i and H_0, respectively. The argument t in $\underline{e}(t)$ and $H_i(t)$ has been dropped for convenience. The fault detection problem reduces to finding the value of i that maximizes equation (7).

To calculate Λ_i the conditional probability densities of $\underline{e}$ under each hypothesis must be found. Since $\underline{e}$ is a jointly Gaussian random vector, it is sufficient to find its mean and covariance matrix, which, from equation (4), are given by

$$\underline{m}_i = E[\underline{e}|H_i] = K_i \underline{f}_i \text{ and } R_i = E[(\underline{e}-\underline{m}_i)(\underline{e}-\underline{m}_i)^T|H_i] = \sigma_a^2 I. \tag{8}$$

Here, $\underline{f}_i := [\underline{f}_i^t(t\text{-}N)\ \underline{f}_i^t(t\text{-}N+1) \ldots \underline{f}_i^t(t)]^T$, the $\underline{f}_i^t(\cdot)$ are as in definition (5), and I is the identity matrix of dimension N+1. Thus, $\underline{m}_i$ is dependent on H_i, but R_i is not. Using equation (8) in the multivariate Gaussian probabality density and substituting into equation (7) gives

$$\Lambda_i = \exp\left\{\frac{1}{2\sigma_a^2}\left[2\underline{e}^T\underline{m}_i - \underline{m}_i^T\underline{m}_i\right]\right\}. \tag{9}$$

The LRT can be further simplified by noting that, since ln(x) is monitonically increasing in x, maximizing Λ_i is equivalent to maximizing $\ln(\Lambda_i)$. Taking the log of equation (9) and defining the statistic

$$S_i := 2\underline{e}^T\underline{m}_i - \underline{m}_i^T\underline{m}_i = 2\underline{e}^T K_i \underline{f}_i - K_i^2 \underline{f}_i^T \underline{f}_i, \tag{10}$$

the LRT becomes:

- choose H_i such that S_i is maximized
- choose H_0 if $S_i < 0 \ \ \forall i \in \{1, 2, \ldots, M\}$ (11)

3.2 The MLE of K_i

The implementation of the LRT described in equations (10) and (11) would require knowing K_i. Since, in practice, such *a priori* knowledge is an unrealistic requirement, the LRT must be modified. A common technique to circumvent this problem is to find the maximum likelihood estimate (MLE) of K_i under H_i, substitute that estimate into equation (10) for K_i, and then use the test given in equation (11). The resulting test is then referred to as a generalized likelihood ratio test (GLRT). LRTs and GLRTs are explained in a general context in Van Trees (1968).

From equation (8) and the definition of the multivatiate Gaussian probability density function, the conditional probability density of $\underline{e}$, given H_i and K_i, is

$$p_{\underline{e}|H_i}(\underline{e}|H_i,K_i) = \frac{1}{\left(2\pi\sigma_a^2\right)^{(N+1)/2}} \exp\left\{\frac{-(\underline{e}-K_i\underline{f}_i)^T(\underline{e}-K_i\underline{f}_i)}{2\sigma_a^2}\right\}. \tag{12}$$

The MLE of K_i under H_i is then

$$\hat{K}_i := \underset{K_i}{\text{argmax}} \left\{ p_{\underline{e}|H_i}(\underline{e}|H_i,K_i)\right\} = \frac{\underline{e}^T\underline{f}_i}{\underline{f}_i^T\underline{f}_i}, \tag{13}$$

which is obtained by setting the partial derivative with respect to K_i equal to zero and solving for K_i.

As will be shown in the following paragraphs, $\hat{K}_i$ has the desirable properties of being both an unbiased and efficient estimate of K_i. That $\hat{K}_i$ is unbiased is easily proven in the following claim.

Claim (1): Under H_i, $\hat{K}_i$ is an unbiased estimate of K_i.

proof: Under H_i the expected value of $\hat{K}_i$ is, from equation (13),

$$E[\hat{K}_i \mid H_i, K_i] = E\left[\frac{\underline{e}^T\underline{f}_i}{\underline{f}_i^T\underline{f}_i} \,\middle|\, H_i, K_i\right] = \frac{K_i\underline{f}_i^T\underline{f}_i}{\underline{f}_i^T\underline{f}_i} = K_i,$$

where the second equality follows from equation (8). ♦

The variance of any unbiased estimate of a nonrandom parameter is always bounded below by what is commonly referred to as the Cramer-Rao bound. If the parameter to be estimated is denoted *a*, the data from which *a* is estimated denoted $\underline{\mathbf{R}}$, and the estimate itself denoted $\hat{a}(\underline{\mathbf{R}})$, then the Cramer-Rao inequality is of the form (Van Trees, 1968)

$$\text{Var}[\hat{a}(\underline{\mathbf{R}})] \geq \left\{-E\left[\frac{\partial^2 \ln p_{\underline{\mathbf{R}}|a}(\underline{\mathbf{R}}|a)}{\partial a^2}\right]\right\}^{-1}, \tag{14}$$

and the quantity to the right of the inequality is referred to as the Cramer-Rao bound. Any estimate which satisfies equation (14) with an equality is called efficient. It is a well known fact (Van Trees, 1968) that an efficient estimate exists if and only if $\frac{\partial \ln p_{\underline{\mathbf{R}}|a}(\underline{\mathbf{R}}|a)}{\partial a}$ can be written in the form $\frac{\partial \ln p_{\underline{\mathbf{R}}|a}(\underline{\mathbf{R}}|a)}{\partial a} = [\hat{a}(\underline{\mathbf{R}}) - a]\psi(a)$, where $\psi(a)$ is any arbitrary function of *a*, but not a function of the data, $\underline{\mathbf{R}}$. Furthermore, if an efficient estimate exists, it must be the maximum likelihood estimate (Van Trees, 1968). We now make use of these facts to prove that $\hat{K}_i$ is effecient.

Claim (2): Under H_i, $\hat{K}_i$ is an efficient estimate of K_i.

proof: Differentiating the logarithm of equation (12) with respect to K_i gives

$$\frac{\partial \ln p_{\underline{e}|H_i,K_i}(\underline{e}|H_i,K_i)}{\partial K_i} = \left[\hat{K}_i - K_i\right]\frac{\underline{f}_i^T\underline{f}_i}{\sigma_a^2}, \tag{15}$$

after substituting the expression for $\hat{K}_i$ from equation (13). Since this satisfies the condition given in the preceeding paragraph, an efficient estimate of K_i exists, and, therefore, the

maximum likelihood estimate, $\hat{K}_i$, must be efficient. ♦

Since $\hat{K}_i$ is efficient, its variance must equal the Cramer-Rao bound of equation (14). Differentiating equation (15) a second time and substituting into equation (14) (with an equality) gives

$$\mathrm{Var}[\hat{K}_i \mid H_i, K_i] = \frac{\sigma_a^2}{\underline{\Gamma}_i^T \underline{\Gamma}_i}. \tag{16}$$

3.3 Extension to a GLRT

Having derived the maximum likelihood estimate of K_i, the LRT can now be extended to the GLRT. Substituting $\hat{K}_i$ of equation (13) for K_i in equation (10) gives

$$S_i = 2\underline{\varepsilon}^T \hat{K}_i \underline{\Gamma}_i - \hat{K}_i^2 \underline{\Gamma}_i^T \underline{\Gamma}_i = \frac{(\underline{\varepsilon}^T \underline{\Gamma}_i)^2}{\underline{\Gamma}_i^T \underline{\Gamma}_i}. \tag{17}$$

From equation (17) it is apparent that $S_i \geq 0 \ \forall i \in \{1, 2, \ldots, M\}$, which is a result of the maximization involved in the maximum likelihood estimation of K_i. Consequently, if the test of equation (11) were used for the GLRT, H_0 would never be chosen. Because of this, one additional modification of the test must be made. A threshold γ must be prescribed so that the test becomes:

- choose H_i such that $S_i = \dfrac{(\underline{\varepsilon}^T \underline{\Gamma}_i)^2}{\underline{\Gamma}_i^T \underline{\Gamma}_i}$ is maximized
- choose H_0 if $S_i < \gamma \ \forall i \in \{1, 2, \ldots, M\}$. (18)

The threshold should be chosen to balance the probability of false alarm and the probability of detection, for which guidelines will be discussed in section 3.4.

The test of equation (18) has a physical interpretation as a correlation receiver. The inner product in the numerator, $\underline{\varepsilon}^T \underline{\Gamma}_i$, represents the "correlation" between the residual vector and the hypothesized fault signature. After scaling that quantity by the "power" of the fault signature, $\underline{\Gamma}_i^T \underline{\Gamma}_i$, the hypothesis whose corresponding fault signature is best correlated with the residual vector is selected. It is in this sense, by correlating the residual vector with the fault signatures, that the dynamics of the fault signature are taken into account and the GLRT achieves improved detection performance.

3.4 Threshold and Window Length Selection

3.2.1 Threshold Selection. As mentioned in the previous section, the threshold γ should be chosen to achieve a suitable balance between the probability of false alarm, denoted α, and the probability of detection, denoted $1-\beta$. To relate γ to α, the distribution of S_i under H_0 must be determined.

Since $K_0 = 0$ under H_0, from equation (8) it is apparent that $\underline{\varepsilon} \sim \mathrm{NID}(\underline{0}, \sigma_a^2 I)$ under H_0, where $\underline{0}$ is a vector of zeros and I is the identity matrix of appropriate dimension. From equation (17),

$$S_i = \frac{(\underline{\varepsilon}^T \underline{\Gamma}_i)^2}{\underline{\Gamma}_i^T \underline{\Gamma}_i} = \frac{(\underline{\varepsilon}^T \underline{\Gamma}_i)(\underline{\varepsilon}^T \underline{\Gamma}_i)}{\underline{\Gamma}_i^T \underline{\Gamma}_i} = \underline{\varepsilon}^T \left[\frac{\underline{\Gamma}_i \underline{\Gamma}_i^T}{\underline{\Gamma}_i^T \underline{\Gamma}_i} \right] \underline{\varepsilon}. \tag{19}$$

Since the matrix in brackets in equation (19) is idempotent, symmetric, and rank one, from the Fisher-Cochran Theorem (Rao, 1973) it follows that under H_0

$$S_i \sim \sigma_a^2 \chi_1^2, \tag{20}$$

where χ_1^2 is a chi-squared random variable with one degree-of-freedom. Equation (20) reveals that under H_0 the distribution of the S_i are identical, independent of i. In other words, given that H_0 is true, each of the alternative hypotheses are equally likely (or unlikely) to be chosen.

From equation (20), at a particular time t and given a particular i and γ, the probability that $S_i \geq \gamma$, given that H_0 is true, can be found. If this probability is denoted α', then, by definition

$$\alpha' = P[S_1 \geq \gamma \mid H_0 \text{ true}] = P[S_2 \geq \gamma \mid H_0 \text{ true}] = \ldots = P[S_M \geq \gamma \mid H_0 \text{ true}]. \tag{21}$$

Since a false alarm occurs if any $S_i \geq \gamma$, the true probability of false alarm α is greater than or equal to α'. Because $\{S_i: i = 1, 2, \ldots, M\}$ are not independent of each other, α is very difficult to calculate analytically. α will, in general, depend on the degree of interdependence of the S_i's, which depends on $G^{-1}(B)$. α may, especially for large M, be significantly greater than α'. In spite of this, the following may serve as a guideline for selecting γ if α' is chosen to be conservatively small. From equations (20) and (21), select γ such that

$$\gamma = \sigma_a^2 \chi_1^2(1 - \alpha'), \tag{22}$$

where $\chi_1^2(b)$ is the b^{th} quantile for a chi-squared random variable with one degree-of-freedom.

This procedure is only a rough guideline for selecting γ based on the probability of false alarm. More work concerning this facet of the test certainly needs to be conducted. The probability of detection depends on the fault signature and fault magnitude also, and is thus even more complicated to determine than the probability of false alarm. Since exact analytical relationships between α, β, and γ are extremely complicated, Monte Carlo techiques may be of some value here.

3.2.2 Window Length Selection. In order to implement the GLRT, the window length N, as well as γ, must be selected. As with γ, an exact analytical relationship between N and the test performance is difficult to evaluate. In selecting N, the tradeoff is between fault detectability and computational expense. The advantage of a large N is that there is a better chance of detecting small magnitude faults, although increasing N will not increase detection speed. The main disadvantage is in the increased computational expense, evident from equation (18).

There are a number of factors to consider when selecting N. Suppose some other statistical test for fault detection, with an out of control ARL of N', is used in conjunction with the GLRT. If the purpose of the GLRT is fast detection of the faults and the other test can consistently detect faults within, say, N' samples, there may be no need to select N larger than N'.

Another factor to consider when selecting N is the dynamics of the fault signature. For the fault detection problem of this paper, a measure of the detectability of a fault occurring at time t-N is $\dfrac{K_i^2 \underline{\Gamma}_i^T \underline{\Gamma}_i}{\sigma_a^2}$ (Van Trees, 1968). Here, the index i is such that $\underline{\Gamma}_i$ is the N+1 length fault signature vector for the unit magnitude fault of interest occurring at time t-N. This detectability measure has a physical interpretation as the square of the Euclidean distance between the mean vector of the residuals under fault and no fault conditions, divided by the variance of the residuals. If, for a particular value of N, $\underline{\Gamma}_i$ is such that good detectability is ensured, there is no need to choose N any larger. In addition, if the system is such that, after a certain point, increasing N does not make the detectability measure any larger, then increasing N further does not provide better detectability. In the simulations of the subsequent sections, the above factors were considered when selecting N, and N = 20 was chosen for both examples.

4 IMPLEMENTATION AND SIMULATION RESULTS

Although in more general settings GLRTs may be complicated to implement, for the process and fault models considered in this paper the implementation is relatively simple. Given that G(B) and σ_a^2 are known, for the hypothesized types and occurrence times of the faults, the various $\underline{\Gamma}_i$ can be calculated off-line. After selection of the threshold, the GLRT requires a bank of M correlation receivers, desribed by equation (18), to be implemented on-line. At each time the residual e(t) is calculated, $\underline{e}(t)$ is updated, and S_i is calculated for i = 1, 2, . . ., M. In addition, $\hat{K}_i$ can be calculated using equation (13). The S_i are then compared to γ and to each other, and the appropriate hypothesis is chosen according to (18).

To illustrate the GLRT method, simulation results are presented in this section. In the simulation a step mean shift and a spike are added (at different times) to a simulated ARMA(2,1) process, and the GLRT method is used to detect both faults. These two types of faults are considered both because they accurately represent many processes and for the sake of simplicity. Given the time of occurrence, both faults are completely described by one parameter, the fault magnitude. The method can be generalized and applied to more complicated faults, such as piecewise step functions or ramps, by using two or more parameters to describe the fault.

Simulation Results

In this example the output data was generated according to equations (1) and (2) using an ARMA(2,1) model with $\Theta(B) = 1-0.5B$, $\Phi(z) = 1-1.8B+0.9B^2$, and $\sigma_a^2 = 1$. Using these values, it can be easily shown that $\sigma_x = 5.83$. Furthermore, the following two process faults were added to the process: i) a mean shift of magnigude $K = \sigma_x = 5.83$ from timesteps 100 to 130, and 2) a spike of magnitude $K = 2\sigma_x = 11.67$ at timestep 200. A window length of N = 20 and a threshold $\gamma = 12$ were used. The original output data, y(t), is shown in Figure 1(a). Due to the high autocorrelation of the process, the faults would be difficult to detect using conventional techniques, and they are hardly discernable from Figure 1(a). The whitened residuals e(t) obtained by passing y(t) through $G^{-1}(B)$ are shown in Figures 1(b) and 1(c). The faults, especially the spike at time 200, is much more apparent in e(t) than in y(t). Figure 1(b) shows the entire sequence of residuals, while Figure 1(c) shows only the residuals during the initial stages of the faults. In Figure 1(c) the solid line represents the fault signature of the true faults, and the dotted line the actual residuals, which follow the fault signature quite closely. Figure 2 shows the results of the fault detection throughout the course of the simulation. Table 1 summarizes the simulation results during the periods in which the faults occurred. At approximately timestep 140 the algorithm incorrectly decided a spike had occurred. The cause of this is that the algorithm tests for step change in the mean from 0 to some nonzero value, and not vice-versa. Consequently, the change in the mean from 5.83 back to 0, not being one of the fault types tested for, was interpreted as a negative spike in the data at timestep 131. This error is not of major concern, however, because the mean shift at timestep 100 had already been detected. After detecting the mean shift, if the new estimate of the mean had been subtracted out of the data, the shift back to 0 in the mean could have been detected just as the original shift was. Barring this period of the simulation, no false alarms occurred. In addition, both faults were detected immediately with no delay. Table 1 shows that the correct fault times were estimated for both the step and spike faults and that the estimated fault magnitudes were close to the true values.

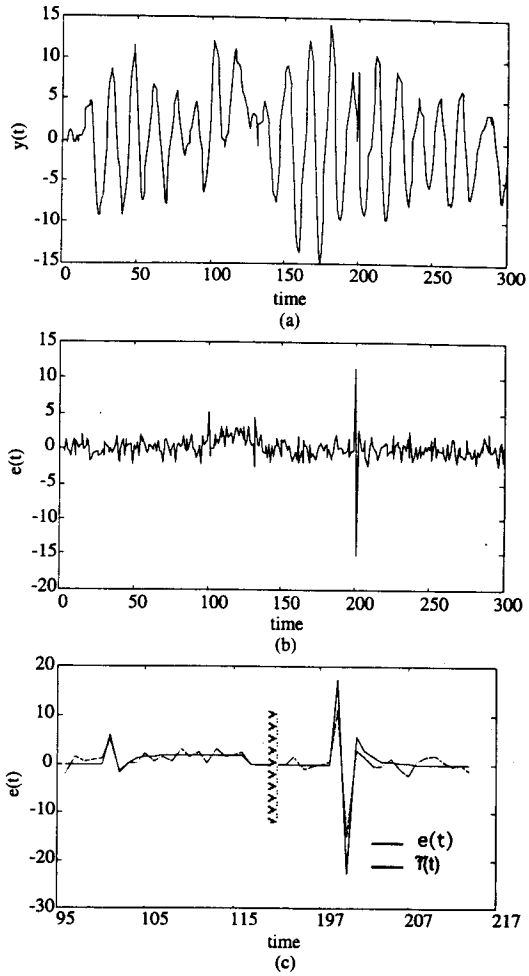

Figure 1 Simulation data: (a) process output with mean shift and spike; (b) whitened residuals; (c) whitened residuals at the onset of the faults and the corresponding fault signatures.

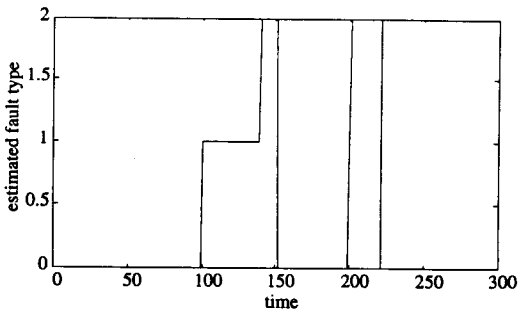

Figure 2 Simulation results for fault detection: 0 - no fault; 1 - step mean shift; 2 - spike.

time	estimated fault type	estimated fault time	estimated magnitude
98	none	-	-
99	none	-	-
100	step	100	5.10
101	step	100	5.01
102	step	100	5.01
103	step	100	4.88
104	step	100	5.09
105	step	100	4.86
106	step	100	4.84
107	step	100	4.65
108	step	100	5.01
109	step	100	5.00
110	step	100	5.15
198	none	-	-
199	none	-	-
200	spike	200	11.47
201	spike	200	11.61
202	spike	200	11.49
203	spike	200	11.48
204	spike	200	11.44
205	spike	200	11.43
206	spike	200	11.44
207	spike	200	11.44
208	spike	200	11.43
209	spike	200	11.43
210	spike	200	11.43

Table 1 Simulation results during the initial stages of the faults. True fault magnitudes were 5.83 (mean shift) and 11.67 (spike).

As previously mentioned, the α and β errors are difficult to evaluate analytically. However, the in control and out of control ARLs for detecting the mean shifts for the above example were investigated using Monte Carlo techniques. For two values of the threshold ($\gamma = 12$ and $\gamma = 15$) and a window length of N = 20, the ARL under no fault conditions and under mean shifts of various magnitudes were estimated using 50 simulations. The results are summarized in Table 2, where K represents the magnitude of the mean shift. In addition to the in control, or no fault, condition (K = 0), three different mean shift magnitudes were considered: $K = \sigma_x$, $K = 0.75\sigma_x$, and $K = 0.5\sigma_x$ (where $\sigma_x = 5.83$). For both $\gamma = 12$ and $\gamma = 15$ the mean shift of magnitude σ_x was detected on the first timestep during all 50 simulations. With no fault present the ARL was 465 for $\gamma = 12$ and 1335 for $\gamma = 15$.

ARL for various mean shift magnitudes				
threshold γ	$K = \sigma_x$	$K = 0.75\sigma_x$	$K = 0.5\sigma_x$	K = 0
$\gamma = 12$	0	0.92	5.47	465
$\gamma = 15$	0	1.58	7.34	1335

Table 2 Monte Carlo results for calculating the ARL for various mean shift magnitudes.

5 SUMMARY AND CONCLUSIONS

For the past two decades GLRTs have been widely used for fault detection in the context of automatic control and monitoring of dynamic systems. This paper suggests that the GLRT has considerable potential for being a valuable tool in quality control also, integrating fault detection with fault estimation and classification.

In this paper a GLRT has been developed to detect faults in processes that are autocorrelated, assuming the process to be of ARMA type driven by white Gaussian noise. The GLRT not only detects the faults, but also classifies the faults according to a pre-specified set of fault types, estimates the time of fault occurrence, and estimates the magnitude of the fault. Furthermore, it has been shown that the estimate of the fault magnitude is both unbiased and efficient, and its variance was derived. The GLRT takes the form of a correlation receiver, correlating the actual residuals with the hypothesized fault signature, where the fault signature is the pattern of the process fault as it appears in the residuals. This has the desirable characteristic of being easy to implement and computationally inexpensive. A complete procedure for implementing the GLRT was developed, including guidelines for selecting the threshold and window length.

Two types of faults were considered in the simulation of this paper: 1) a step change in the process mean, and 2) a spike. The simulation demonstrates the effectiveness of the GLRT method, and a Monte Carlo simulation is included in an analysis of the ARL properties for the case of a mean shift. The GLRT's effectiveness results from the fact that it takes into account the dynamics of the fault signature, improving detection performance considerably. Work is now being conducted for extending the GLRT to more complicated types of faults, such as exponentially drifting means and linearly drifting means, for which the complexity of the test should increase only moderately.

References

Alt, F. B., Deutsch, S. J. and Walker, J. W., 1977, "Control Charts for Multivariate, Correlated Observations," *ASQC Technical Conference Transactions*, Philadelphia, Pa.

Basseville, M., 1988, "Detecting Changes in Signals and Systems - A Survey," *Automatica*, Vol. 24, No. 3, pp. 309-326.

Box, G. E. P. and Jenkins, G. M., 1976, *Time Series Analysis, Forecasting, and Control*, Holden Day, Oakland, Ca.

Deshayes, J. and Picard, D., 1986, "Off-line Statistical Analysis of Change-point Models Using Non Parametric and Likelihood Methods," In Basseville, M. and Benveniste, A. (eds), *Detection of Abrupt Changes in Signals and Dynamical Systems*, Chapter 5, pp. 103-168, LNCIS No. 77, Springer, Berlin.

Dooley, K. J. and Kapoor, S. G., 1990, "An Enhanced Quality Evaluation System for Continuous Manufacturing Processes, (Part 1: Theory; Part 2: Application)," *ASME Journal of Engineering for Industry*, Vol. 112, pp. 57-68.

Dooley, K. J., Kapoor, S. G., Dessouky, M. I. and Devor, R. E., 1986, "An Integrated Quality Systems Approach to Quality and Productivity Improvement in Continuous Manufacturing Processes," *ASME Journal of Engineering for Industry*, Vol. 108, pp. 322-327.

Johnson, R. A. and Bagshaw, M., 1974, "The effect of Statistical Correlation on the Performance of CUSUM tests," *Technometrics*, Vol. 16, No. 1, pp. 103-112.

Montgomery, D. C. and Mastrangelo, C. M., 1991, "Some Statistical Process Control Methods for Autocorrelated Data," *Journal of Quality Technology*, Vol. 23, No. 3, pp. 179-193.

Notohardjono, B. D. and Ermer, D.S., 1986, "Time Series Control Charts for Correlated and Contaminated Data," *ASME Journal of Engineering for Industry*, Vol. 108, pp. 219-226.

Pandit, S. M. and Wu, S. M., 1990, *Time Series and System Analysis with Applications*, Wiley, N. Y.

Rao, C. R., 1973, *Linear Statistical Inference and Its Applications, 2nd edition*, Wiley, N. Y.

Tsay, R. S., 1988, "Outliers, Level Shifts, and Variance Changes in Time Series," *Journal of Forecasting*, Vol. 7, pp. 1-20.

Van Trees, H. L., 1968, *Detection, Estimation, and Modulation Theory, Part I*, Wiley, N. Y.

Vasilopoulos, A. V. and Stamboulis, A. P., 1978, "Modification of Control Chart Limits in the Presence of Data Correlation," *Journal of Quality Technology*, Vol. 10, No. 1, pp. 20-30.

Willsky, A. S., 1976, "A Survey of Design Methods for Failure Detection in Dynamic Systems," *Automatica*, Vol. 12, pp. 601-611.

Willsky, A. S. and Jones, H. L., 1976, "A Generalized Likelihood Ratio Approach to the Detection and Estimation of Jumps in Linear Systems," *IEEE Transactions on Automatic Control*, Vol AC-21, No. 1, pp. 108-112.

Yashchin, E., 1993, "Performance of CUSUM Control Schemes for Serially Correlated Observations," *Technometrics*, Vol. 35, No. 1, pp. 37-52.