Identifying Change Point in Production Time-Series Volatility Using Control Charts and Stochastic Differential Equations

MARTIN KOVARIK
Tomas Bata University in Zlin, Faculty of Management and Economics
Department of Statistics and Quantitative Methods
Mostni 5139, 760 01 Zlin
CZECH REPUBLIC
m1kovarik@fame.utb.cz

Abstract: The article focuses on volatility change point detection using SPC (Statistical Process Control) methods, specifically time-series control charts and stochastic differential equations (SDEs). Contribution will review recent advances in change point detection for the volatility component of a process satisfying stochastic differential equation (SDE) based on discrete observations, and also by using time-series control charts. Theoretical part will discuss methodology of time-series control charts and SDEs driven by a Brownian motion. Research part will demonstrate the methodologies in a simulation study focusing on analysis of the AR(1) process by means of time-series control charts and SDEs. The aim is to make use of change point detection in time series of production processes and highlight versatility of control charts not only in manufacturing but also in managing financial cash flow stability.

Key–Words: Autocorrelation, Change point, Control charts, Outliers, Stochastic Differential Equations (SDE), Average Run Length (ARL), Statistical Process Control (SPC)

1 Introduction

Traditional Statistical Process Control (SPC) methods, e.g. Shewhart's and CUSUM control charts assume data to be independent. However, the assumption has been challenged when data were found to be serially correlated in many real-world applications. With autocorrelation, performance of the traditional methods is considerably reduced. Majority of the studies focus on mean shift detections while variability shift detections in autocorrelated structures of the time series are omitted due to special causes affecting system's variability.

This paper will focus on change point problem for process volatility via stochastic differential equation with observations collected at discrete intervals. The instant the change occurred in volatility regime is identified retrospectively by maximum likelihood method on approximated likelihood. For continuous-time observations of diffusion processes, Lee, Nishiyama and Yoshida (2006) considered change point estimation for the drift. We will only assume regularity conditions in the drift process.

Contribution's aim to compare behavior of $\lambda_{LS,max}$ and $\bar{\lambda}_{LS}$ control charts by Atienza et al. (1998) [2] and utilized for detecting AR(1) process shift level, an approach to detect time-series volatility change point by stochastic differential equations for discrete processes. Their performance will be evaluated and compared with SCC (Special Causes Control modified Shewhart's control chart for residuals) by Alwan and Roberts and Monte Carlo simulation in the R Programming Language's spc library [32] will be used as a comparison tool.

2 Literature Review

The article first provides brief overview of works pertaining to time-series volatility change point identification in SPC of manufacturing companies. The topic is actively researched as autocorrelation and mean shifts in time series when analyzing automatically-collected data are frequent.

Author has studied change point detection using parametric and non-parametric procedures extensively. In some cases, studies were carried out for known underlying distributions, e.g., binomial, Poisson, Gaussian and normal. The chapter discusses selected works on change point detection.

Traditional SPCs assume data to be independent. However, the assumption has been challenged when data were found to be serially correlated in many real-world applications. With autocorrelation, performance of the traditional methods is considerably reduced. This prompted work by Alwan and Roberts

(1988) [1] who proposed expected monitoring error after selecting suitable time series model for a given process. The method is intuitive in case autocorrelation is explained in the model, and residuals are process-independent random errors; it can therefore be used to monitor residuals. Literature dealing with SPC is divided into two categories: those based on time-series models, and those independent on them. For the former, three approaches were devised: monitoring residuals, direct observations, and using novel statistical characteristics their brief outline is provided further on. Wardell et al. (1994) [29] and Lua and Reynolds (1999) [23] exponentially-weighted moving averages (EWMA) to monitor residuals. Castagliola and Tsung (2005) [3] investigated influence of non-normality on control charts and created modified Shewhart control chart for residuals special causes control (SCC) which is more robust when non-normality is encountered.

CUSUM is a widely used change point detection algorithm. Basseville and Nikiforov (1993) [11] described four different derivations. The first is intuition-based and uses ideas connected to simple signals integrations with adaptive thresholds; the second is based on repeated use of sequential probability ration test; the third uses off-line point of view for multiple hypotheses testing; the fourth is based on open ended tests. The principle of CUSUM stems from stochastic hypothesis testing method (Chen et al., 2005 in [8]).

Nazario et al. (1997) [12] developed a sequential test procedure for transient detections in a stochastic process which can be expressed as an autoregressive moving average (ARMA) model. Preliminary analysis shows that if an ARMA(p, q) time series exhibits transient behavior, its residuals behave as an ARMA(Q, Q) process, with $Q \leq p + q$. They further showed residuals in the model before parameter change behave approximately as a sequence of independent random variables; afterwards they become correlated. Based on this fact, Nazario et al. derived a new sequential test to determine when the transient behavior occurs in a given ARMA time series [33].

Blazek et al. (2001) [13] developed efficient adaptive sequential and batch-sequential methods for early detection denial-of-service attacks. Both algorithms used thresholding of test statistics to achieve a fixed rate of false alarms and are based on change point detection theory: detecting changes in statistical models as soon as possible while controlling rate of false alarms. There are three attractive features to the approach: first, both methods are self-learning, enabling adaptation to varying network loads and usage patterns; second, they allow detecting attacks with small average delays for a set false alarm rate; third, they are computationally simple, and hence can be implemented online [34].

Lund et al. (2007) [14] looked at change point detection in periodic and autocorrelated time series using classic change point tests based on sums of squared errors. The method was successfully applied in the analysis of climate changes.

Moskvina and Zhigljavsky (2003) in [15] developed an algorithm based on sequential application of singular-spectrum analysis (SSA) whose main idea is to perform singular value decomposition (SVD) of a trajectory matrix obtained from the original time series and subsequently reconstructing it.

Mboup et al. (2008) in [16] presented a method based on direct online estimation of signal's singularity points. Using piecewise local polynomial representation, the problem is transformed into delay estimation. A change point instant is a solution of a polynomial equation coefficients of which are composed by short-time window iterated integrals of the noisy signal. The method showed good robustness to various types of noises.

Auret and Aldrich (2010) [17] used random forest models to detect change points in dynamic systems; Wei et al. (2010) in [18] Lyapunov exponent and change point detection theory for anomaly detection; Aldrich and Jemwa, (2007) in [19] phase methods to detect changes in complex process systems.

Vincent (1998) in [20] presented a technique for identifying inhomogeneities in Canadian temperature series by applying linear regression models to determine whether the series is homogeneous. Vincent's procedure is a type of "forward regression" algorithm in that the significance of non-change point parameters in the regression model is assessed before (and after) a possible change point is introduced. The most parsimonious model is selected to describe the data and then used to generate residuals which are tested for autocorrelation to determine whether there are inhomogeneities in the series. At first, it considers the entire period and then divides the series into homogeneous segments. Each is defined by some change points and every one corresponds to either an abrupt change in mean level or a change in the trend [35].

3 Problem Formulation

Here, article present mathematical background for detecting time-series volatility shifts using control charts and stochastic differential equations for discrete processes, later utilized to compare ARL performance and sensitivity to mean shifts. Research part will demonstrate the methodologies in a simulation study focusing on analysis of the AR(1) process by means of time-series control charts and SDEs. Their performance will be evaluated and compared with SCC (Special Causes Control modified Shewhart's control chart for residuals) by Alwan and Roberts and Monte Carlo simulation in the R Programming Language's spc library [32] will be used as a comparison tool.

3.1 Change point detection using SDEs

Change point estimation identifies the instant in which a change occurs in model's parameter. Several approaches are possible: we consider a least squares solution other such as maximum likelihood change point estimation also exist. Assume a diffusion process (*Note: For continuous-time observations the problem was studied in [9]. Bayesian approach for discrete-time observations can be found in [27]*). solution to

$$dX_t = b\left(X_t\right)dt + \theta\sigma\left(X_t\right)dW_t, \quad (1)$$

where $b\left(\cdot\right)$ and $\sigma\left(\cdot\right)$ are known functions and $\theta \in \Theta \subset \mathbb{R}$ is the parameter of interest. As in [27], given discrete observations from (1) on $[0, T = n\Delta_n]$, we want to retrospectively identify if and when a change in θ occurred and consistently estimate it before and after the change point. The asymptotes is $\Delta_n \to 0$ as $n \to \infty$ and $n\Delta_n = T$ fixed. (*Note: For ergodic diffusion processes and $n\Delta_n = T \to \infty$, the results are valid under additional mild regularity conditions.*) To simplify, assume the change occurs at k_0, one of the integers in $1, \ldots, n$. This is a case of volatility change point estimation frequently occurring in financial applications. Assume $\theta = \theta_1$ before the time change and afterwards with $\theta_1 < \theta_2$ (irrelevant for final results). To obtain a simple least squares estimator, Euler approximation is used under the assumption all necessary requirements have been met. It is of the form (Hinkley, 1971 in [4]).

$$X_{i+1} = X_i + b\left(X_i\right)\Delta_n + \theta\sigma\left(X_i\right)\left(W_{i+1} - W_i\right)$$

and standardized residuals

$$Z_i = \frac{\left(X_{i+1} - X_i\right) - b\left(X_i\right)\Delta_n}{\sqrt{\Delta_n}\sigma\left(X_i\right)} = \theta\frac{\left(W_{i+1} - W_i\right)}{\sqrt{\Delta_n}}.$$

Z_i's are i.i.d. (independent and identically distributed) Gaussian random variables. Change point estimator is obtained from (Inclan and Tiao, 1994 in [5])

$$\hat{k}_0 = \arg\min_k \left(\min_{\theta_1,\theta_2}\left\{\sum_{i=1}^{k}\left(Z_i^2 - \theta_1^2\right)^2 + \right.\right. \\ \left.\left. + \sum_{i=k+1}^{n}\left(Z_i^2 - \theta_2^2\right)^2\right\}\right) \quad (2)$$

with, $k = 2, \ldots, n-1$. $[x]$ will denote the integer part of real x, $k_0 = [n\tau_0]$ and $k = [n\tau]$, $\tau, \tau_0 \in (0,1)$ indicate change point on a continuous timescale. Partial sums are defined as $S_n = \sum\limits_{i=1}^{n} Z_i^2$, $S_k = \sum\limits_{i=1}^{k} Z_i^2$, $S_{n-k} = \sum\limits_{i=k+1}^{n} Z_i^2$, $\bar{\theta}_1^2$ and $\bar{\theta}_2^2$ denote least squares estimators of θ_1^2 and θ_2^2 for a given k in (2),

$$\bar{\theta}_1^2 = \frac{S_k}{k} = \frac{1}{k}\sum_{i=1}^{k} Z_i^2,$$

and

$$\bar{\theta}_2^2 = \frac{S_{n-k}}{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n} Z_i^2.$$

They will be refined once a consistent estimator of k_0 is found. U_k^2 denotes the quantity

$$U_k^2 = \sum_{i=1}^{k}\left(Z_i^2 - \bar{\theta}_1^2\right)^2 + \sum_{i=k+1}^{n}\left(Z_i^2 - \bar{\theta}_2^2\right)^2.$$

So $\hat{k}_0$ is then $\hat{k}_0 = \arg\min\limits_k U_k^2$.

To study the asymptotic properties of U_k^2, it is better to rewrite it: $U_k^2 = \sum\limits_{i=1}^{n}\left(Z_i^2 - \bar{Z}_n\right)^2 - nV_k^2$, where $\bar{Z}_n = \frac{1}{n}\sum\limits_{i=1}^{n} Z_i^2$ and

$$V_k = \left(\frac{k\left(n-k\right)}{n^2}\right)^{\frac{1}{2}}\left(\bar{\theta}_2^2 - \bar{\theta}_1^2\right) = \frac{S_n D_k}{\sqrt{k\left(n-k\right)}}$$

with $D_k = \frac{k}{n} - \frac{S_k}{S_n}$.

U_k^2 is obtained by lengthy but straightforward algebra. It is useful because minimizing U_k^2 is equivalent to maximizing V_k and hence D_k. The following estimator is there for easier to consider (Lee, Nishiyama and Yoshida, 2006 in [10]):

$$\hat{k}_0 = \arg\max_k |D_k| = \arg\max_k \left(k\left(n-k\right)\right)^{\frac{1}{2}}|V_k|. \quad (3)$$

A side remark: for fixed k (and under suitable hypotheses), D_k is approximate likelihood ratio statistic for testing the null hypothesis of no change in volatility. Once $\hat{k}_0$ has been obtained the following estimators of θ_1 and θ_2 can be used (Csörgö and Horvath, 1997 in [24]):

$$\hat{\theta}_1^2 = \frac{S_{\hat{k}_0}}{\hat{k}_0} \quad (4)$$

$$\hat{\theta}_2^2 = \frac{S_{n-\hat{k}_0}}{n-\hat{k}_0} \quad (5)$$

Results provide consistency of $\hat{k}_0$, $\hat{\theta}_1^2$ and $\hat{\theta}_2^2$ as well as their asymptotic distributions.

Lemma 1 *In Stanton's approach, they are Nadaraya-Watson kernel regression estimators of the following conditional expectations (Kutoyants, 1994 in [30])*

$$b\left(x\right)=\lim_{t\to 0}\frac{1}{t}\mathrm{E}\left\{X_t-x|X_0=x\right\}$$

$$\sigma^2\left(x\right)=\lim_{t\to 0}\frac{1}{t}\mathrm{E}\left\{\left(X_t-x\right)^2|X_0=x\right\}.$$

$b\left(x\right)$ and $\sigma^2\left(x\right)$ are seen as instantaneous conditional means and variances when $X_0=x$ in the process. For fixed Δ_n the two quantities are rewritten as (Kutoyants, 2004 in [31])

$$b\left(x\right)=\frac{1}{\Delta_n}\mathrm{E}\left\{X_{i+1}-X_i|X_i=x\right\}+\frac{o\left(\Delta_n\right)}{\Delta_n},$$

$$\sigma^2\left(x\right)=\frac{1}{\Delta_n}\mathrm{E}\left\{\left(X_{i+1}-X_i\right)^2|X_i=x\right\}+\frac{o\left(\Delta_n\right)}{\Delta_n}.$$

If Z_i's have been estimated in this case, the following contrast to identify the change point and be used (Liechty and Roberts, 2001 in [25]), (Iacus, 2008 in [27]) and (Iacus and Yoshida, 2009 in [28]):

$$\begin{aligned}\tilde{k}_0=\arg\min_k\Bigg\{&\sum_{i=1}^{k}\left(\hat{Z}_i^2-\tfrac{\hat{S}_k}{k}\right)^2+\\&+\sum_{i=k+1}^{n}\left(\hat{Z}_i^2-\tfrac{\hat{S}_{n-k}}{n-k}\right)^2\Bigg\},\end{aligned}\tag{6}$$

where $\hat{S}_k=\sum_{i=1}^{k}\hat{Z}_i^2$ and $\hat{S}_{n-k}=\sum_{i=k+1}^{n}\hat{Z}_i^2$. We have a new statistic

$$\hat{V}_k=\left(\frac{k\left(n-k\right)}{n^2}\right)^{\frac{1}{2}}\left(\frac{\hat{S}_{n-k}}{n-k}-\frac{\hat{S}_k}{k}\right)=\frac{\hat{S}_n\hat{D}_k}{\sqrt{k\left(n-k\right)}},$$

where $\hat{D}_k=\frac{k}{n}-\frac{\hat{S}_k}{\hat{S}_n}$.

The change point is identified as a solution of $\hat{k}_0=\arg\max_k\left|\hat{D}_k\right|$. Consistency and distributional results provided in (De Gregorio and Iacus, 2007 in [26]).

This paper will focus on change point problem for process volatility via stochastic differential equation with observations collected at discrete intervals. The instant the change occurred in volatility regime is identified retrospectively by maximum likelihood method on approximated likelihood. For continuous-time observations of diffusion processes, Lee, Nishiyama and Yoshida (2006) considered change point estimation for the drift. We will only assume regularity conditions in the drift process.

3.2 Detecting outliers and mean shift levels using control charts based on time-series models

Consider an ARMA model of the form

$$\phi\left(B\right)Z_t=\phi_0+\theta\left(B\right)\varepsilon_t\tag{7}$$

where Z_i is a stationary time series representing a measurement process,

$$\phi\left(B\right)=1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p$$

is autoregressive polynomial of order p,

$$\theta\left(B\right)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q$$

a moving average polynomial of order q, B a backshift operator, and $\{\varepsilon_t\}$ a sequence of independent and normally-distributed random errors with zero mean and constant variance σ^2. Without loss of generality, $\{Z_t\}$'s ϕ_0 level will be fixed at zero. Let $\hat{Z}_t$ be a predicted value from a suitably-selected ARMA model. Then residuals (Chandra, 2001 in [21])

$$\begin{aligned}\Big\{&e_1=Z_1-\hat{Z}_1,e_2=Z_2-\hat{Z}_2,\ldots,e_t=\\&=Z_t-\hat{Z}_t,\ldots\Big\}\end{aligned},$$

where Y_t and $f\left(t\right)$ represent "contaminated" time series with non-standard exogenous disturbances, outliers and different time series shift levels. Depending on the disturbance, $f\left(t\right)$ can be either deterministic or stochastic. For the former, $f\left(t\right)$ is of the form

$$f\left(t\right)=\omega_0\frac{\omega\left(B\right)}{\delta\left(B\right)}\xi_t^{(d)},\tag{8}$$

where

$$\xi_t^{(d)}=\begin{cases}1, & when\ t=d,\\ 0, & when\ t\neq d,\end{cases}\tag{9}$$

is a variable expressing whether the disturbance occurred at time d, $\omega\left(B\right)$ and $\delta\left(B\right)$ backshift polynomials describing disturbance's dynamic effect on Y_t, and ω_0 a constant denoting disturbance's initial influence (Harris and Ross, 1991 in [22]).

If $\omega\left(B\right)/\delta\left(B\right)=1$, the disturbance is an additive outlier. AO (*Note: Additive outliers affect single observation. After the disturbance, the time series returns to normal. Data coding error is additive outlier*) affects time-series' mean only if $t=d$. The root cause is often a data coding error. In discrete production processes, AO may occur when mixed information from different base materials is gathered. If $\omega\left(B\right)/\delta\left(B\right)=\theta\left(B\right)/\phi\left(B\right)$, equation (9) represents innovative outliers influencing Y_t up to $t=d$

after which the effect recedes exponentially. In continuous chemical processes, IO (*Note:Innovative outliers are compounded with noise at a particular time-series point. In a stationary time series, they affect several observations, in a non-stationary ones, every observation from an initial point*) is very likely caused by contamination. Parameters will be strongly influenced in $t = d$ during preventive maintenance, e.g., when a component (conduit etc.) is replaced with a contaminated one. After $t = d$, the contamination effect recedes. If $\omega(B)/\delta(B) = 1/(1-B)$, the disturbance in equation (9) constitutes level shift, changing Y_t' beginning in $t = d$ upwards or downwards which lasts until $t > d$. LS (*Note: Level shift outliers behave like step functions, i.e., affecting every observation from an initial point in the series by a constant. For a stationary process, the shift changes its mean after the initial point, and the process turns into a non-stationary one*) is primarily caused by change in material or process setup quality. We use ω_{AO}, ω_{IO} and ω_{LS} to denote whether ω_0 is tied to AO, IO or LS. From equations (8) and (9):

$$\frac{\phi(B)}{\theta(B)} Y_t = \frac{\phi(B)\,\omega(B)}{\theta(B)\,\delta(B)} \xi_t^{(d)} \omega_0 + \varepsilon_t.$$

If $y_t = [\phi(B)/\theta(B)]\, Y_t$ and

$$x_t = [\phi(B)\,\omega(B)/\theta(B)\,\delta(B)]\, \xi_t^{(d)},$$

$y_t = \omega_0 x_t + \varepsilon_t$ a simple regression model. ω_0 is estimated using (Bai, 1994 in [6]) $\hat{\omega}_0 = \frac{\sum_{t=1}^{T} y_t x_t}{\sum_{t=1}^{T} x_t^2}$ with $\mathrm{Var}(\hat{\omega}_0) = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2}$, where T is sample size. Using the equations, the following ω_0 estimates for the three disturbance types described above are obtained:

$$\hat{\omega}_{AO,t} = \begin{cases} \rho_{AO,t}^2 \left(y_t - \sum_{i=1}^{T-t} \pi_i y_{t+i} \right) & t = 1, 2, \ldots, T-1 \\ y_t & t = T, \end{cases} \tag{10}$$

$$\hat{\omega}_{IO,t} = y_t, t = 1, 2, \ldots, T, \tag{11}$$

$$\hat{\omega}_{LS,t} = \begin{cases} \rho_{LS,t}^2 \left(y_t - \sum_{i=1}^{T-t} \eta_i y_{t+i} \right) & t = 1, 2, \ldots, T-1 \\ y_t & t = T, \end{cases} \tag{12}$$

where $\rho_{AO,t}^2 = \left(1 + \sum_{i=1}^{T-t} \pi_i^2\right)^{-1}$, $\rho_{IO,t}^2 = 1$, $\rho_{LS,t}^2 = \left(1 + \sum_{i=1}^{T-t} \eta_i^2\right)^{-1}$, π_i and η_i represent B^i coefficients in polynomials

$$\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots = \phi(B)/\theta(B)$$

and

$$\eta(B) = 1 - \eta_1 B - \eta_2 B^2 - \ldots = \pi(B)/(1-B).$$

Weights π are computed by multiplying both sides of $\pi(B)$, and $\theta(B)$ to get $\theta(B)\left(1 - \pi_1 B - \pi_2 B^2 - \ldots\right) = \phi(B)$. For ARMA(1, 1):

$$\begin{aligned} 1 - \phi B &= (1 - \theta B)\left(1 - \pi_1 B - \pi_2 B^2 - \ldots\right) = \\ &= 1 - (\pi_1 + \theta) B - (\pi_2 - \theta\pi_1) B^2 - \\ &- (\pi_3 - \theta\pi_2) B^3 - \ldots \end{aligned}$$

Comparing the coefficients such as powers of B, $\pi_1 = \phi - \theta, \pi_2 = \theta\pi_1$ and $\pi_j = \theta\pi_{j-1} = \theta^{j-1}\pi_1$ for $j > 1$ are calculated. Similar approach can be utilized for weights η. For ARMA(1, 1), they are $\eta_1 = \phi - \theta - 1$ and $\eta_j = \eta_{j-1} + \theta^{j-1}\pi_1$ for $j > 1$. For AR(1), it holds that $\pi_j = \phi^j$ and $\eta_j = \phi - 1$ for $j \geq 1$ (Kim, Cho and Lee, 2000 in [7]). A characteristic can therefore be constructed for testing the existence of AO, IO and LS in time d:

$$\lambda_{j,d} = \frac{\hat{\omega}_{j,d}}{\left[\mathrm{Var}(\hat{\omega}_{j,d})\right]^{1/2}} = \frac{\hat{\omega}_{j,d}}{\rho_{j,d}\sigma}, \tag{13}$$

where $j = AO, IO, LS$.

Under the null hypothesis of outliers and level shifts not present, and for known d and ARMA model parameters in equation (7), $\lambda_{AO,t}, \lambda_{IO,t}$ and $\lambda_{LS,t}$ are sampled from asymptotic distribution, e.g., N(0,1). In practice, the parameters are usually unknown and need to be replaced by consistent estimates (Lee, Ha, O. Na and S. Na, 2003 in [8]). To detect AO, IO or LS at an unspecified position we will compute the following statistic:

$$\lambda_{j,\max} = \max_{1 \leq t \leq T} \{|\lambda_{j,t}|\}, \tag{14}$$

where $j = AO, IO, LS$.

The null hypothesis of AO, IO or LS not present is rejected when $\lambda_{j,max}$ is higher than a critical value. However, sampling distribution in equation (14) is challenging to determine exactly. Chen and Liu (1993) considered broad spectrum of π weights in percentile estimation with the help of Monte Carlo simulation, and noted the percentiles tied to equation (12) are substantially lower than those tied to equations (10) and (11). For instance, when T = 200, they claimed the first percentile estimation for $\lambda_{LS,max}$ lies between 3.3 and 3.5 while for $\lambda_{AO,max}$ and $\lambda_{IO,max}$ it lies between 4.0 and 4.2 in AR(1).

SPC focuses on detecting presence of exogenous variance sources exhibited by changes in a process level or variance. If the change is caused by AO, IO or

LS, the root cause can be narrowed down, facilitating quick response and correction of problems linked to changed signal of a process change.

Next, article will show how SPC process monitoring system based on detecting outlier statistics and changes in its level can be created.

4 Problem Solution

Control charts $\lambda_{LS,max}$ and $\bar{\lambda}_{LS}$ by Atienza et al. (1998) utilized for process level shifts detection in AR(1) will be compared to volatility changes detection using stochastic differential equations for discrete processes. Their performance will be compared with SCC by Alwan and Roberts (special causes control chart modified Shewhart's control chart for residuals). Monte Carlo simulations in the R Programming Language's spc library will be used as a comparison tool.

5 Monitoring System

Earlier mathematical background suggests a regulation procedure for process level shifts detection either by monitoring $\lambda_{LS,max}$ or

$$\bar{\lambda}_{LS,t} = \frac{\sum_{t=1}^{T} \lambda_{LS,t}}{T}.$$

This article focus on LS in AR(1) when devising the control charts, and plot shift estimates obtained from SDE based on volatility change-point estimator for diffusion processes based on least squares to the graph of the original time series. As the drift coefficient is unknown, $dX_t = b\,(X_t)\,dt + \Theta \cdot dW_t$ is considered and b is estimated non-parametrically. We start with a set of m observations from a controlled process (i.e., statistically managed). The first phase sees it very likely affected by outliers and level shifts whose presence may lead to substandard model identification and incorrect assessment of process deviations. Methods allowing outlier detection and model parameter estimation should thus be preferred.

The analysis shows $\lambda_{LS,t}$ control charts to be sensitive to correctly selecting m, e.g., m = 100 gives incorrect information about Average Run Length (ARL (*Note:Average Run Length (ARL) denotes mean number of steps before the statistic crosses a regulation band, or before CL change (shift) is detected*)). Satisfactory ARL is achieved with $m = 200$, usually corresponding to at most several seconds or minutes in real-time process monitoring for SPC, even when m is high.

When the initial set of observations is at available, $\lambda_{LS,t}$ can be computed. Using the regulation apparatus we only need to preserve m newest observations, i.e., when a new observation is added, the oldest one is discarded. With m points, m values for $\lambda_{LS,t}$ is calculated. Suppose we start with a set $\{Z_1, Z_2, \ldots, Z_m\}$ and m respective values for $\lambda_{LS,t}$, $\{\lambda_{LS,1}, \lambda_{LS,2}, \ldots, \lambda_{LS,m}\}$. When Z_{m+1} is supplied, $\{Z_2, Z_3, \ldots, Z_{m+1}\}$ scale is used to calculate $\{\lambda_{LS,2}, \lambda_{LS,3}, \ldots, \lambda_{LS,m+1}\}$. Starting from the initial set m, after arbitrary number of steps i we have $\{Z_{i+1}, Z_{i+2}, \ldots, Z_{m+i}\}$ from which $\{\lambda_{LS,i+1}, \lambda_{LS,i+2}, \ldots, \lambda_{LS,m+i}\}$ is computed. Mathematically, we follow

$$\lambda_{LS,\max,i} = \max_{i+1 \le t \le m+i} \{|\lambda_{LS,t}|\},$$

with $i = 1, 2, 3, \ldots$, or

$$\bar{\lambda}_{LS,i} = \frac{\sum_{t=i+1}^{m+i} \lambda_{LS,t}}{m}, \quad i = 1, 2, 3, \ldots$$

Graph indicates the process is not statistically managed, $\lambda_{LS,max}$ or $\bar{\lambda}_{LS,t}$ crosses its respective regulation band set based on $\lambda_{LS,max}$ and $\bar{\lambda}_{LS,t}$ sampling distributions of which are challenging to quantify due the statistics being functions of dependent variables. It is thus recommended to simulate and input standard or acceptable ARL under the assumption the process is statistically managed. To determine regulation bands to produce the ARL value in a statistically-managed state, a software tool should be run. Starting points for simulations in the $\lambda_{LS,max}$ diagram are reproducible from Chen and Liu (1993).

5.1 Performance, characteristics and behavior of control charts

For AR(1) with $\phi > 0$, expected residual value at and $t = d$ is ω_{LS} when a level shift occurs in range ω_{LS}. For $t > d$, the value is $(1 - \phi)\,\omega_{LS}$. In case of SCC, it is highly probable the shift will be detected as soon as it occurs; afterwards, the probability decreases rapidly, especially when $\phi \to 1$. SCC is expected to perform better than CUSUM when the shift in $t = d$ generates high residuals (e.g., more than 3 in absolute expression). However, if SCC fails to detect the occurrence, CUSUM is recommended instead. Both have their advantages and disadvantages for monitoring autocorrelated processes.

The $\lambda_{LS,t}$ control chart combines desirable properties of SCC and CUSUM: it is capable to detect sudden and small changes in the process level. To demonstrate, two LS scenarios for AR(1) with $\phi = 0.9$ and

$\sigma = 1$ were selected. The former is depicted in Figure 1; here, the change did not produce high residuals. First 200 observations are from AR(1) with $\phi = 0.9$ and zero mean, the shift occurs at $t = 201$. Three control charts were constructed: SCC, CUSUM on residuals and $\lambda_{LS,max}$. SDE was used in the original time series. Parameters for the ARL were selected in statistically-managed state to be approximately 370. SCC (Figure 2B) failed to identify the change in $201 \leq t \leq 400$ interval; CUSUM, $\lambda_{LS,max}$ and SDE detected it almost at the same time, the latter at $t = 208$ with parameters $\Theta_1 = 574952.9$ and $\Theta_2 = 63678.35$.

The second scenario analyzes $\lambda_{LS,max}$ for LS producing high residuals. Time series in Figure 3A originate from identical process as in Figure 2A. Compared to previous demonstration, LS in $t = 201$ generates high residuals, causing SCC to signal "out of regulation bounds". CUSUM on residuals failed in the task. Figure 3D shows $\lambda_{LS,max}$ managed to detect the shift in $t = 201$, as did SDE. Examples in Figures 2 and 3 prove $\lambda_{LS,max}$ $(t = 202)$ and SDE $(t = 201$ with $\Theta_1 = 363861.2$ and $\Theta_2 = 54765.15)$ achieved better results than SCC and CUSUM on residuals.

5.2 Comparison of performance and behavior using ARL

For different ϕ and δ, performance of $\lambda_{LS,max}$ and $\bar{\lambda}_{LS}$, SCC, and SDE was compared using Monte Carlo simulation. However, shifts in δ is measured by standard deviation. Every run consisted of 5,000 iterations, the algorithm was written in the R Programming Language. AR(1) was simulated using the spc package, sde package was used to load stochastic differential equations. $m = 200$, regulation bands for $\lambda_{LS,max}$ and $\bar{\lambda}_{LS}$ were set so that ARL in a statistically-managed state approximately corresponded to ARL in SCC with $\pm 3\sigma$.

Figure 1 shows dominance of control charts based on $\lambda_{LS,max}$, and SDE further exacerbated when $\phi > 0$. This is expected as $\lambda_{LS,max}$ integrates benefits of SCC and CUSUM on residuals. Compared to $\lambda_{LS,max}$, the $\bar{\lambda}_{LS}$ diagram performs better when detecting small shifts but is less sensitive for big shifts, especially when ϕ is high.

6 Discussion

Traditional control charting procedures are based on the assumption process observations are i.i.d. With advent of high-speed data collection, the assumption is usually violated, i.e., autocorrelation among measurements which causes significant deterioration in

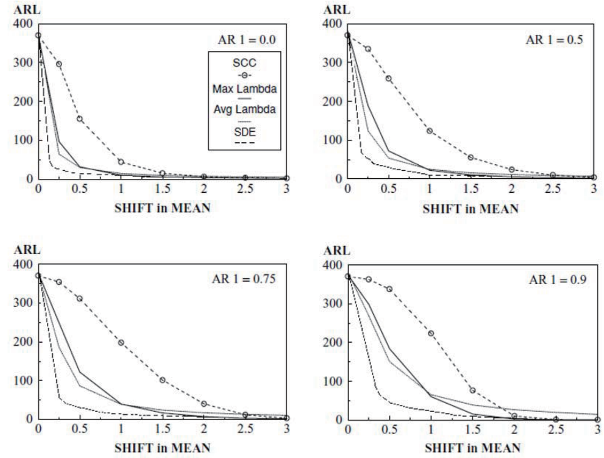

Figure 1: ARL comparison: SCC, $\lambda_{LS,max}$, $\bar{\lambda}_{LS}$ and SDE. (Source: own work, R)

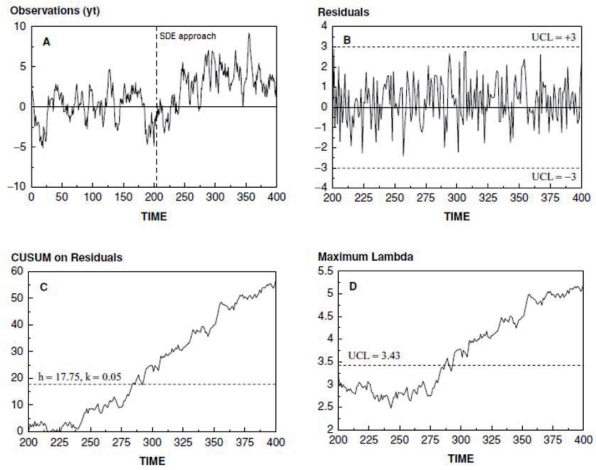

Figure 2: Performance of SCC, SUCUM on residuals, $\lambda_{LS,max}$ and SDE when detecting level shifts for low residuals at the change point. (Source: own work, R)

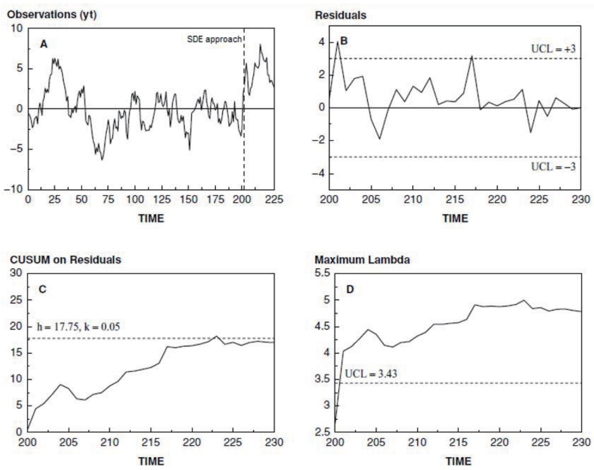

Figure 3: Performance of SCC, SUCUM on residuals, $\lambda_{LS,max}$ and SDE for detecting level shifts for high residuals at the change point. (Source: own work, R)

control charting performance becomes inherent for a stable process. Several methods for handling autocorrelated processes have been proposed, the most popular utilizes Shewhart's, CUSUM or EWMA charts of appropriately-fitted ARMA model. However, they exhibit poor sensitivity, particularly for positively autocorrelated processes. As an alternative, we have explored statistics used in a time-series procedures for outlier and level shift detection. The study focused on level shifts in autocorrelated processes with emphasis on the AR(1) model. The results showed time-series charts are found to be sensitive for detecting small shifts and we exploited the fact these control charts can be used in certain situations where data are autocorrelated.

As for the sub-series identified by the change point estimate and estimators of their parameters: both are consistent and asymptotically normal at the are of $\sqrt{n}$ with n the number of observations. Least squares estimator $\hat{\tau}_0$ seems to have good performance in terms of bias and variability for models with constant or bounded drifts while it fares badly in presence of unbounded drift as time T grows.

7 Conclusion

Most traditional control charts are based on the assumptions process observations are independent and sampled from identical distributions. With proliferation of high-speed automatic data collection, the former is usually violated, i.e., autocorrelation becomes inherent characteristic of a stable process and substantial performance degradation follows. Several ways to address the problem have been devised, the most popular utilizing Shewhart's, CUSUM or EWMA control charts on residuals of approximately-fitted ARMA model. They nevertheless have low sensitivity, especially when positively-autocorrelated processes are involved. As an alternative, we investigated statistics used in time-series procedures for outlier and mean shift detection.

The study focused on mean level shifts for autocorrelated processes with emphasis on the AR(1) model. The results proved the apparatus for monitoring the changes can be based on SDE, $\lambda_{LS,max}$ or $\bar{\lambda}_{LS}$. $\lambda_{LS,max}$ combines beneficial properties of Shewhart's and CUSUM diagrams, exhibiting superior ARL performance in comparison with other methods. Unlike SCC and $\lambda_{LS,max}$, $\bar{\lambda}_{LS}$ is more sensitive to small shifts, and less sensitive to large shifts. The proposed regulation apparatus is expandable to identify additive and innovative outliers. Intervention type identification affecting the process allows observing sources of statistical unmanageability, an important step in special causes of variance elimination. It can be also used for more general autoregressive integrated moving averages (ARIMA).

Stochastic differential equations rank among the most widely-used stochastic models to describe continuous financial time series. Although data are collected at discrete intervals, model structure enables detailed data analysis. Analyzing AR(1), it revealed (unlike time-series control charts) another variability change point and appears to be preferable for change point detection. SDEs also provide robustness of the estimates.

The time series model based approach is easy to understand and effective in some situations. However, it requires identifying appropriate time series model from a set of initial in-control data. This may not be easy to establish in practice and may be too complicated to practicing engineers. Hence, the model-free approach has recently attracted much attention.

Such autocorrelation causes significant deterioration in control charting performance. In order to address this, several approaches for handling autocorrelated processes have been proposed, the most popular one utilizing either Shewhart, CUSUM or EWMA chart of the residuals of the appropriately fitted ARMA model. However, procedures of this type demonstrate poor sensitivity, especially when dealing with positively autocorrelated processes.

Uncorrelated observations appear when automatically collecting data, usually by software upgradeable to include SPC for data processing. In such a system, serviceability of the procedure can be optimized.

Acknowledgements: The author is also thankful to the Operational Programme Education for Competitiveness co-funded by the European Social Fund (ESF) and national budget of the Czech Republic for the grant No. CZ.1.07/2.3.00/20.0147 - "Human Resources Development in the field of Measurement and Management of Companies, Clusters and Regions Performance", which provided financial support for this research.

References

[1] L. C. Alvan, and H. V. Roberts, Time-series modelling for statistical process control, *Journal of Business and Economic Statistics,* vol. 16, 1988, pp. 87-95.

[2] O. O. Atienza, L. C. Tang and B. V. Ang, A SPC procedure for detecting level shifts of autocorrelated processes, *Journal of Quality Technology,* vol. 30, 1988, pp. 340-351.

[3] P. Castagliola, and F. Tsung, Autocorrelated SPC for non-normal situations, *Quality and Reli-*

ability Engineering International, vol. 21, 2005, pp. 131-161.

[4] D. V. Hinkley, Inference about the change-point from cumulative sum tests, *Biometrika,* vol. 58, 1971, pp. 509-523.

[5] C. Inclan and G. C. Tiao, Use of Cumulative Sums of Squares for Retrospective Detection of Change of Variance, *Journal of the American Statistical Association,* vol. 89, 1994, pp. 913-923.

[6] J. Bai, Least Squares Estimation of a Shift in Linear Processes, *Journal of Times Series Analysis,* vol. 15, 1994, pp. 453-472.

[7] S. Kim, S. Cho and S. Lee, On the Cusum Test for Parameter Changes in GARCH(1,1) Models, *Communications in Statistics - Theory and Methods,* vol. 29, 2000, pp. 445-462.

[8] S. Lee, J. Ha, O. Na and S. Na, The Cusum Test for Parameter Change in Time Series Models, *Scandinavian Journal of Statistics,* vol. 30, 2003, pp. 781-796.

[9] G. Chen, Y. K. Choi and Y. Zhou, Nonparametric Estimation of Structural Change Points in Volatility Models for Time Series, *Journal of Econometrics,* vol. 126, 2005, pp. 79-144.

[10] S. Lee, Y. Nishiyama and N. Yoshida, Test for Parameter Change in Diffusion Processes by Cusum Statistics Based on One-Step Estimators, *Annals of the Institute of Statistical Mathematics,* vol. 58, 2006, pp. 211-222.

[11] M. Basseville and I. Nikiforov, *Detection of abrupt changes: theory and application,*Englewood Cliffs, Ed. New York: Prentice Hall, 1993, ch. 2.

[12] D. Nazario, B. Ramirez and S. Tep, Transient detection with an application to a chemical process, *Computers ind. Eng,* 1997, pp. 896-908.

[13] R. B. Blazek, K. HongJoong, R. Boris and T. Alexander, A novel approach to detection of "denial of service" attacks via adaptive sequential and batch-sequential change-point detection methods, *in Proceedings of the 2001 IEEE,* 2001, pp. 220-226.

[14] R. Lund, L. W. Xiaolan, Q. Lu, J. Reeves,, C. Gallagher and Y. Feng, Change-point detection in periodic and autocorrelated time series, *Journal of Climate,* vol. 20, 2007, pp. 5178-5190.

[15] V. Moskvina and A. Zhigljavsky, An algorithm based on singular spectrum analysis for change-point detection, *Communication in Statistics, Simulation and Computation,* vol. 32, 2003, pp. 319-352.

[16] M. Mboup, C. Join and Ml. Fliess, An on-line change-point detection method, *in Proceedings of the Mediterranean Conference on Control and Automation Congress Centre,* Ajaccio, France: IEEE, 2008, pp. 1290-1295.

[17] L. Auret and C. Aldrich, Unsupervised process fault detection with random forests, *Industrial and Engineering Chemistry Research,* vol. 49 (19), 2010, pp. 9184-9194.

[18] X. Wei, H. Hanping, Y. Yue and Q. Wang, Anomaly detection of network traffic based on the largest Lyapunov exponent, *IEEE,* 2010, pp. 581-585.

[19] C. Aldrich and G. T. Jemwa, Detecting change in complex process systems with phase space methods, *The 2007 Spring National Meeting, 2007-AIChE,* 2007, pp. 1-10.

[20] L. A. Vincent, A technique for the identification of inhomogeneities in Canadian temperature series, *Journal of Climate Change,* 1998, pp. 1094-1104.

[21] M. J. Chandra, Statistical Quality Control. *United States of America: CRC Press, LLC,* 2001, ch. 8.

[22] T. J. Harris and W. H. Ross, Statistical process control procedures for correlated observations, *Canadian Journal of Chemical Engineering,* vol. 69, 1991, pp. 48-57.

[23] C. W. Lu and M. R. Reynolds, Control charts for monitoring the mean and variance of autocorrelated processes, *Journal of Quality Technology,* vol. 31, no. 3, 1999, pp. 259-274.

[24] M. Csörgö and L. Horvath, Limit Theorems in Change-point Analysis. *New York: Wiley,* 1997, ch. 1.

[25] J. Liechty and G. Roberts, Markov chain Monte Carlo methods for switching diffusion models, *Biometrika,* vol. 88, no. 2, 2001, pp. 299-315.

[26] A. De Gregorio and S. M. Iacus (2007). Least squares volatility change point estimation for partially observed diffusion processes, Working Paper. Available: http://arxiv.org/abs/0709.2967v1.

[27] S. M. Iacus, Simulation and Inference for Stochastic Differential Equations: with R Examples, *Springer Series in Statistics, New York: Springer,* 2008, ch. 4.

[28] S. M. Iacus and N. Yoshida (2007). Estimation for the Change Point of the Volatility in a Stochastic Differential Equation. Available: http://arxiv.org/abs/0906.3108.

[29] D. G. Wardell, H. Moskowitz and R. D. Plante, Run-length distributions of specialcause control charts for correlated processes (with discussion). *Technometrics,* vol. 36, 1994, pp. 3-27.

[30] Y. Kutoyants, *Identification of Dynamical Systems with Small Noise*, Dordrecht: Kluwer, 1994, ch. 5.

[31] Y. Kutoyants, *Statistical Inference for Ergodic Diffusion Processes*. London: Springer-Verlag, 2004, ch. 4.

[32] YUIMA Project Team. (2010). Yuima: The YUIMA Project Package. Available: http://R-Forge.R-project.org/projects/yuima.

[33] C. Ting and Ch. Su, Using Concepts of Control Charts to Establish the Index of Evaluation for Test Quality, *WSEAS Transactions on Communications,* vol. 8, 2009, pp. 535-544.

[34] N. K. Huat and H. Midi, Robust Individuals Control Chart for Change Point Model, *WSEAS Transactions on Mathematics,* vol. 9, 2010, pp. 7-16.

[35] A. S. Jamali and L. JinLin, False Alarm Rates for the Shewhart Control Chart with Interpretation Rules, *WSEAS Transactions on Information Science and Applications,* vol. 3, 2006, pp. 7–10.