Performance of control charts for autoregressive conditional heteroscedastic processes

YUE FANG¹ & JOHN ZHANG², ¹Lundquist College of Business, University of Oregon, USA and ²Department of Mathematics, Indiana University of Pennsylvania, USA

ABSTRACT  This paper examines the robustness of control schemes to data conditional heteroscedasticity. Overall, the results show that the control schemes which do not account for heteroscedasticity fail in providing reliable information on the status of the process. Consequently, incorrect conclusions will be drawn by applying these procedures in the presence of data conditional heteroscedasticity. Control charts with time-varying control limits are shown to be useful in that context.

1 Introduction

Traditional control charts, such as the Shewhart, cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) control schemes, have been widely used in statistical process control (SPC). They serve as on-line SPC procedures to monitor process stability, to detect assignable variation, or to forecast process movements in industrial processes and other applications. A typical control chart is constructed from a sample collected when the process is in control. To determine the parameters of control charts, assumptions about the data generated by the process have to be made.

In standard SPC applications, a state of control is identified with a process generating independent and identically distributed (iid) normal random variables. In practice, it is often difficult to attain a state of control in this strict sense. In situations where the normality assumption is violated or the independence assumption is not satisfied, the control charts based on the iid assumptions may be
ineffective and inappropriate. It is well known that distributions with heavier tails may increase the presence of outliers in the data. However, in general, the control charts are not sensitive to the normality assumption and work reasonably well unless the population is markedly non-normal.

If data are not independent, then the control charts which do not take account of autocorrelation could wrongly infer that the process is in control or signal that an assignable cause might have occurred. More sophisticated control schemes have been proposed for serially correlated observations based on the autoregressive moving average (ARMA) models (see, for example, Alwan & Roberts, 1988; Harris & Ross, 1991; Wardell et al., 1994). ARMA (Box & Jenkins, 1994) processes are linear models. If the disturbance term is assumed to be normally distributed, then the analysis of ARMA processes is within the Gaussian framework. We can regard the process based on ARMA models as ‘in control’ in a broader sense—a sense that goes beyond the simple benchmark of iid random variables.

In order to obtain a better understanding of the process, it is often required to modify the standard SPC procedures so that the model assumptions are approximately satisfied and a rigorous analysis is possible. For example, serial correlation, seasonality, missing values and non-constant variance are some common features in environmental, biological and chemical data (Berthouex et al., 1978). In order to apply SPC charts to those data, a wider class of stochastic models may be required.

In this paper, we extend the study of SPC procedures to some non-linear processes. More specifically, we investigate the robustness of control charts to data conditional heteroscedasticity. In some applications, data conditional heteroscedasticity is common and is often associated with processes that have heavy tails and variability clusters. For example, Cuthbertson and Gasparro (1993) studied manufacturing inventories and found that the general autoregressive conditional heteroscedastic (GARCH) model is consistent with existing theories and with UK data. Weiss (1984) analyzed 16 US economic series from the Citibank Economic Database and fitted ARCH-type models to the data with varying degrees of success.

Heteroscedasticity has an important consequence for control problems. In this paper, we provide evidence that the control schemes, which do not account for heteroscedasticity, are not robust to data conditional heteroscedasticity. Our study indicates that, in general, the conventional calculated control limits are invalid. The in-control average run length (ARL) falls substantially and the magnitude depends on the parameters of conditional variance processes. We also develop a simple control scheme with time-varying control limits and show that the proposed procedure is useful in dealing with conditional heteroscedasticity data.

To facilitate the study, we assume that the processes can be described by either GARCH or ARMA–ARCH processes. GARCH models represent one type of non-linear model, while the conditional variance varies over time; ARMA–ARCH processes are ARMA models with ARCH errors (see De Gooijer and Kumar (1992) for a review of non-linear models, and Tong (1990) for further discussion on the subject). The non-linearity stems from the conditional heteroscedasticity of the disturbance term. GARCH and ARMA–ARCH processes have been successfully applied in many time series (see Bollerslev et al. (1992) for a survey). There are various reasons for the presence of conditional heteroscedasticity in the observed sequence. One reason is that any changes in the timescale of data sampling may create ARCH effects in the observed time series (Stock, 1988).

Among a wide variety of control schemes, the Shewhart, CUSUM and EWMA
schemes are basic and popular. All three charts have been constructed under the normality and independence assumptions. They share some common appealing aspects, such as being easy to set up, implement and interpret. We will focus on the EWMA chart. The conclusions for the Shewhart and CUSUM charts are similar and the results are not reported in this paper.

This paper is organized as follows. Section 2 describes the GARCH (1,1) model, which serves as a data generator. In Section 3, the performance of the EWMA control scheme is evaluated in the presence of data conditional heteroscedasticity. The ARLs, as a measurement of performance, are assessed through a simulation study, given that a single step shift in the process mean occurs. An adjusted control scheme with time-varying control limits subject to data conditional heteroscedasticity is discussed in Section 4. Section 5 extends our discussion to the control schemes designed for serially correlated data. Concluding remarks are presented in Section 6.

2 GARCH (1,1) model

A wide variety of models based on the ARCH type were developed by Engle (1982) and were generalized by Bollerslev (1986). In this section, we consider the GARCH(1,1) model. The GARCH(1,1) model considers a stochastic process $y_t$ in equidistant discrete time $t$. The distribution of observations $\{y_t\}$ is specified by the distribution of $y_t$ conditional on its past values $y_{t-1}$. Without loss of generality, the process mean, when the process is in control, is assumed to be zero. The model can be formulated as

\[ y_t = \sigma_t \varepsilon_t \] (1)

and

\[ \sigma_t^2 = \gamma + \alpha \sigma_{t-1}^2 + \beta \sigma_{t-1}^2 \] (2)

where the parameters $\gamma, \alpha, \beta \geq 0$, and $\alpha + \beta < 1$. $\{\varepsilon_t\}$ are iid standard normal. The parameters $\alpha$ and $\beta$ are particularly interesting, because they provide information about the persistence of temporal shocks.

The conditional distribution of $y_t$ is normal with mean zero and standard deviation $\sigma_t$. However, the unconditional distribution of $y_t$ is not of any standard form. By the usual ARMA analog, the process is weakly stationary with mean zero and constant unconditional variance

\[ \sigma^2 = \text{Var}(y_t) = \frac{\gamma}{1 - \alpha - \beta} \] (3)

The existence of higher moments of $y_t$ depends on the levels of both $\alpha$ and $\beta$. The distribution of $y_t$ is symmetric, since all the odd moments of $y_t$ are zero. The kurtosis of $y_t$, i.e. $\kappa$, is given by

\[ \kappa = \frac{E[(y_t - E(y_t))^4]}{\text{Var}^2(y_t)} = \frac{3 - 3(\alpha + \beta)^2}{1 - 3\alpha^2 - \beta^2 - 2\alpha\beta} \] (4)

The kurtosis exists only for $(1 - 3\alpha^2 - \beta^2 - 2\alpha\beta) > 0$ with $\alpha, \beta \geq 0$. In equation (4), $\kappa$ in (4) is greater than 3 if either $\alpha$ or $\beta$ is greater than zero; therefore, the model yields observations with heavier tails than those of a normal distribution.
One can verify equation (4) by applying the fact that $y_t^2$ is ARMA(1,1), i.e.

$$y_t^2 = \gamma + (\alpha + \beta) y_{t-1}^2 + a_t - \beta a_{t-1}$$

where $a_t = \sigma_t^2 (\omega_t - 1)$ is white noise (Harvey, 1993). In the interests of brevity, we do not include the detailed computation here. Figure 1 shows the regions of $\alpha$ and $\beta$ in which the variance and the kurtosis of $y_t$ exist.

3 EWMA scheme in the presence of data conditional heteroscedasticity

The general model of control charts consists of the center line $\mu$, which is the equilibrium level of some quality characteristic of interest; the upper control limits (UCL) and the lower control limits (LCL), which have the same distance from the center line, expressed in standard deviation units of the observed sequence $\sigma_y$. We have

$$\text{Control limits} = \mu \pm k L \sigma_z,$$

where $k$ is a predetermined constant and the value of the parameter $L$ depends on the design of the chart. In the EWMA case, $z_t = \lambda y_t + (1 - \lambda) z_{t-1}$ and

$$L = \left\{ \frac{\lambda}{2 - \lambda} \left[ 1 - (1 - \lambda)^2 \right] \right\}^{1/2}$$

where $\lambda$ is the smoothing parameter, which is the weight assigned to the last observation. We will have that $\mu = 0$ in equation (5), since the process mean is assumed to be zero when the process is in control.

Theoretical derivations of the ARLs for GARCH processes are difficult. We perform simulation experiments for ARLs when the process mean has shifted by $\rho \sigma$, for some non-negative parameter $\rho$. If $\rho = 0$, then the process is in control. Data are generated by the GARCH(1,1) model defined in equations (1) and (2). The performance of the EWMA scheme is evaluated on the parameter space spanned by $\alpha$ and $\beta$. The levels of $\alpha$ and $\beta$ are in the interval [0.0, 1.0) with increment 0.1. Imposing the weakly stationary condition $\alpha + \beta < 1$ implies that
ARLs are calculated only on half the $\alpha$ and $\beta$ space. We assume that $\sigma^2 = 1$ for all simulations; therefore, the parameter $\gamma = 1 - \alpha - \beta$ from equation (3).

To assess the robustness of the EWMA scheme to the data conditional heteroscedasticity, we perform simulation experiments for the smoothing parameter $\lambda = 0.1$ and 0.3. The multiplier $k$ in equation (5) is chosen so that ARL $= 370$ in the case of no shift in the process mean and no data conditional heteroscedasticity. For example, when $\lambda = 0.1$, $k = 2.715$. If $\lambda = 0.3$, then $k$ increases to be approximately 2.928. More values of $k$ for different in-control ARLs can be found in Lucas and Saccucci (1990). The process is initialized at the in-control value and the process mean shifts immediately after the first observation. All simulations are based on 10,000 replications.

Figures 2–5 report the simulation results for $\lambda = 0.1$ when $\rho = 0.0, 0.5, 1.0$ and

![Contour plot of the ARL for the in-control EWMA with $\lambda = 0.1$.](image1)

![Contour plot for the EWMA with $\lambda = 0.1$, when the process has shifted by half a standard deviation.](image2)
The ARL for an in-control EWMA ($\rho = 0.0$) is shown in Fig. 2. The ARL is less than 370 for most of the $\alpha$ and $\beta$ combinations, except for the regions with $\alpha \approx 0$ or $\alpha + \beta \approx 1$. A significant reduction in ARL occurs if $\alpha$ is not close to zero or $\alpha + \beta$ is not close to unity. The lowest ARL is about 230, which is equivalent to a reduction of 37.84%. In general, the values of the ARL depend on both $\alpha$ and $\beta$. However, $\alpha$ has more impact on ARLs than does $\beta$.

The kurtosis of $y_i$, i.e. $\kappa$, plays an important role and the curve $1 - 3\alpha^2 - \beta^2 - 2\alpha\beta = 0$ divides the $\alpha$ and $\beta$ space into two non-overlapping areas. If $1 - 3\alpha^2 - \beta^2 - 2\alpha\beta > 0$, then $\kappa$ exists. ARL = 370 when the observations are iid normal ($\alpha = \beta = 0$). As $\alpha$ and $\beta$ increase, $\kappa > 3$. Consequently, the ARL decreases, as a result of the conditional heteroscedasticity effects.

The behavior of the ARL is interesting when $1 - 3\alpha^2 - \beta^2 - 2\alpha\beta \leq 0$. In this case, the kurtosis of $y_i$ does not exist. However, the ARL does not continue to drop as
\( \alpha \) and \( \beta \) increase; instead, rather surprisingly, it increases as \( \alpha + \beta \) approaches unity. The lowest ARLs settle around the curve of \( 1 - 3\alpha^2 - \beta^2 - 2\alpha\beta = 0 \). For example, if \( \beta = 0 \), then the kurtosis of \( y \) is given by

\[
\kappa = \begin{cases} 
 3, & \alpha = 0 \\
 3(1 - \alpha^2)/(1 - 3\alpha^2), & \alpha < (1/3)^{1/2} \\
 3, & \alpha > (1/3)^{1/2} \\
 \text{does not exist}, & \alpha = (1/3)^{1/2}
\end{cases}
\]

The ARL first decreases and then increases when \( \alpha \) varies from zero to unity. The lowest ARL is at about 222.7 when \( \alpha \) is around \((1/3)^{1/2} \approx 0.577\).

In order to understand the behavior of the ARL when \( 1 - 3\alpha^2 - \beta^2 - 2\alpha\beta \leq 0 \), let us consider an extreme case where \( \alpha + \beta \approx 1 \). When we set \( \alpha + \beta = 1 \), the GARCH (1,1) model in equations (1) and (2) becomes an integrated GARCH (IGARCH) model. The IGARCH process is no longer weakly stationary, since it does not have a finite second movement. However, the IGARCH process is strictly stationary and has the strange property that, no matter what the starting point, \( \sigma_t^2 \) collapses to zero almost definitely (Nelson, 1990). Hence, the observed series effectively disappears. As \( \alpha + \beta \) approaches unity, the observed series behaves more like an IGARCH process. The conditional heteroscedasticity effects decrease and the ARL climbs back to and even beyond the nominal level (370) for some values of \( \alpha \) and \( \beta \).

Figures 3–5 show the effects on the ARLs when the process mean has shifted. Again, the ARLs are more likely influenced by \( \alpha \) than by \( \beta \). Surprisingly, we find that the ARLs have a strong tendency to increase as \( \alpha \) increases, especially in the cases in which \( \rho = 1.0 \) and \( \rho = 3.0 \). The magnitude of the increment of the ARLs depends on the level of the shift parameter \( \rho \).

Figures 6–9 report the simulation results for \( \lambda = 0.3 \) when \( \rho = 0.0, 0.5, 1.0 \) and 3.0 respectively. The ARL for an in-control EWMA (\( \rho = 0.0 \)) is shown in Fig. 6 and indicates a very similar pattern as for the case \( \lambda = 0.1 \) when \( \alpha \) and \( \beta \) vary. The ARLs are less than 370 for the majority of \( \alpha \) and \( \beta \). The reduction in the ARLs is significant when \( \alpha \) is between 0.2 and 0.8.

Figures 7–9 show the effects on the ARLs when the process mean has shifted.

![Contour plot of the ARL for the in-control EWMA with \( \lambda = 0.3 \).](image)
Unlike the results for $\lambda = 0.1$, the ARLs for $\rho = 0.5$ and $\rho = 1.0$ are not monotonic increasing functions of $\alpha$. Instead, they depend on the levels of both $\alpha$ and $\beta$. The monotonic increasing phenomenon of the ARL that we observed when $\lambda = 0.1$ arises only for $\rho = 3.0$.

In summary, the EWMA s which do not account for heteroscedasticity are generally not robust to data conditional heteroscedasticity. The fourth moment of $y_t$, i.e. $\kappa$, plays an important role in the determination of the ARLs. Overall, the level of $\alpha$ has more impact on the ARLs than does the level of $\beta$. When the process is in control, the EWMA s with both $\lambda = 0.1$ and $\lambda = 0.3$ offer higher out-of-control false alarm rates than those for the data without conditional heteroscedasticity. When the process is in control, the ARLs are in the range 150–370 for most of $\alpha$.
and $\beta$. The reduction in the ARL is more significant for $\lambda = 0.3$ than is that for $\lambda = 0.1$. However, when the process is out of control, the EWMA's with $\lambda = 0.1$ are less sensitive to mean shifts of all three levels compared with the results for the data without conditional heteroscedasticity. In contrast to the case of $\lambda = 0.1$, when the process is out of control, in general, the control scheme with $\lambda = 0.3$ is more sensitive to a small or a median mean shift and is less sensitive to a large mean shift compared with the result for data without conditional heteroscedasticity.

4 Control charts with time-varying control limits

Although the GARCH model defined in equations (1) and (2) implies that the unconditional variance of $\{y_t\}$ is a constant, the conditional variance of $\{y_t\}$ could change over time. The model for the temporal dependence in conditional second moments suggests that time-varying control limits should be appropriate. The set-up in equation (2) allows for using past information to construct the conditional variance of the process as an alternative to constant control limit schemes.

We suggest that the control limits for the EWMA scheme have the format

$$
\text{Control limits} = \mu \pm k\sigma_{z_i}
$$

(7)

where $\sigma_{z_i}^2$ is the conditional variance of $z_i = \lambda y_t + (1 - \lambda)z_{t-1}$ and is given by

$$
\sigma_{z_i}^2 = \lambda^2 \sum_{j=0}^{t-1} (1 - \lambda)^2j \sigma_{i-j}^2 + (1 - \lambda)^2t \sigma_0^2
$$

(8)

The sequence $\{\sigma_i^2\}$ depends on $\{y_t\}$ and the initial value $\sigma_0^2$. Its dynamic is determined by equation (2). We take $\sigma_0^2$ to be unity, so that the unconditional variance of $y$ is unity, as we did in the previous section.

To illustrate control schemes with time-varying control limits, we use a sequence of simulated observations generated by equations (1) and (2) with parameters $a = 0.4$ and $\beta = 0.0$. Some 100 $z_i$ terms are calculated and plotted with two different
control schemes (see Fig. 10). One scheme has constant control limits according to equation (5) and the other scheme uses time-varying control limits based on equations (7) and (8). The control scheme with time-varying control limits utilizes the available information and constantly upgrades the estimation on the process variability. The control scheme with time-varying control limits is insensitive to data conditional heteroscedasticity, so creates fewer false out-of-control alarms.

To evaluate the performance of the proposed control schemes with time-varying control limits, we compute ARLs according to the control limits determined by equations (7) and (8), as well as ARLs based on a 'two-in-a-row' rule for comparison. The two-in-a-row control procedure is used to detect problems in control with heavier-tailed observations than normal. Two successive observations in a row which are outside the limits are considered to represent an out-of-control signal. The two-in-a-row procedure is insensitive to excess variability caused by occasional outliers (Lucas & Saccucci, 1990). Since $\beta$ has less influence on ARLs than does $\alpha$, $\beta$ is taken to be zero in order to simplify the evaluation. The parameter $\alpha$ is set equal to 0.1, 0.2, $\ldots$, 0.9. We only report the results for the EWMA with $\lambda = 0.1$. Similar conclusions can be made for EWMA s with other $\lambda$ values. Again, the control limits are based on a zero-state in-control ARL = 370, as in Section 3.

Figures 11–13 report the ARLs for EWMA s with time-varying control limits and schemes based on the two-in-a-row control rule. For a moderate or large shift of the process mean ($\rho = 1.0$ or 3.0), the scheme with time-varying control limits outperforms the charts based on the two-in-a-row rule. For a small deviation of the process mean ($\rho = 0.5$), the performances of the two control schemes are about the same, except that the scheme with time-varying control limits performs poorly as $\alpha$ approaches unity.
Overall, the control scheme with time-varying control limits is found to perform better than does the scheme based on the two-in-a-row rule, except for the case where \( \alpha \) is close to unity and \( \rho \) is relatively small. Other control schemes based on high moments, such as kurtosis, may be instructive and useful in some applications. However, more sophisticated procedures may be difficult to interpret and implement.
5 Control schemes designed for serially correlated observations

In this section, we extend the discussion to the case of serially correlated observation sequences. The presence of serial correlation in the observation sequence may cause excessive variability and, consequently, the traditional control charts, such as the Shewhart, CUSUM or EWMA schemes, become meaningless. One approach to monitor the serial correlated data is to fit the data to a structure model and examine the residuals or the one-step-ahead forecast error, instead of the original data sequence.

To model both serial correlation and data heteroscedasticity, the combined ARMA–ARCH models are useful. ARMA–ARCH models are a natural extension of ARMA models which allow \( y_t \) to have time-varying conditional second moments. For example, ARMA(1,1)–ARCH(1) can be written as

\[
y_t - \phi y_{t-1} = (1 - \phi) \mu + \varepsilon_t - \theta \varepsilon_{t-1}
\]

where

\[
\sigma_t^2 = E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) = \gamma + \alpha \sigma_{t-1}^2
\]

The process defined by equations (9) and (10) has mean \( \mu \) and variance

\[
\text{Var}(y_t) = \frac{(1 + \theta^2 - 2\phi \theta) \gamma}{1 - \phi^2 - \alpha(1 + \theta^2 - 2\phi \theta)}
\]

As the case without data conditional heteroscedasticity, one approach to deal with both serial correlation and ARCH error structure is to fit the observation sequence to ARMA models and then analyze the residuals by applying control schemes with time-varying control limits as discussed in Section 4.
6 Conclusions

Our study indicates that control schemes which do not account for heteroscedasticity are not robust to data heteroscedasticity. In general, the in-control ARLs fall substantially if data conditional heteroscedasticity is present, and the magnitude depends on the kurtosis of the observed process. The control scheme with time-varying control limits is introduced and the performance is evaluated against the alternative scheme based on the two-in-a-row rule. The Monte Carlo evidence suggests that the control scheme with time-varying control limits works well, making it a useful tool in the presence of heteroscedasticity.

ARCH-type models are non-linear models which include as special cases linear processes such as white noise and ARMA process. Maximum-likelihood estimation procedures can be applied to the GARCH and ARMA–ARCH models, and the algorithm developed by Berndt et al. (1974) provides a convenient method of computation. More details may be found, for example, in Engle (1982), Weiss (1984, 1986) and Bollerslev (1986). Although we have selected ARCH-type models to illustrate SPC procedures in the non-linear settings, in practice, the specific application will dictate which linear and/or non-linear model might be appropriate. The idea proposed in this paper can be extended to deal with other forms of non-linearity. It becomes clear that the model chosen to fit the data is important and any model misspecification will result in false signals of the state of the process. That topic is beyond the scope of the discussion of this paper and readers interested in a comprehensive treatment of it are referred to the literature, such as Weiss (1986), Bollerslev et al. (1993) and Tiao and Tsay (1994).

Acknowledgements

The authors thank Frederick W. Morgan, Gopal K. Kanji (the editor) and the referee for helpful comments and suggestions.

REFERENCES


