ROBUST MONITORING OF CONTAMINATED MULTIVARIATE DATA

Eric B. Howington
Department of Management
Valdosta State University
1500 North Patterson St.
Valdosta, GA 31698-0076
229-245-3876
ebhowington@valdosta.edu

ABSTRACT

Monitoring a process that suffers from data contamination using a traditional multivariate $T^2$ chart can lead to an excessive number of false alarms. This paper extends the diagnostic statistic technique of Davis and Adams [1] to the multivariate case. A traditional $T^2$ control chart augmented by a diagnostic statistic improves the work stoppage rates for contaminated multivariate data and maintains the ability to detect process shifts.

INTRODUCTION

Davis and Adams [1] consider the problem of dealing with contaminated data in univariate control charts. They consider monitoring a process for which measurement systems are problematic, leading to occasional unusual measurements for key quality characteristics. These atypical measurements do not reflect the true state of the process and are referred to as outliers. A sample containing an outlier is said to be contaminated. Contaminated data can be troublesome for practitioners monitoring a process because a control chart signal could indicate a true process shift or could simply be the result of an outlier. Thus, Davis and Adams distinguish two types of signals: signals that indicate a process problem and signals that reflect a data problem. They propose use of a diagnostic statistic that allows the practitioner to distinguish between the two types of signals. When the control chart signals, a diagnostic statistic is calculated for that sample. If the value of the diagnostic statistic exceeds a threshold, then the signal could have been caused by contaminated data and further investigation is warranted before stopping the process. If the value of the diagnostic statistic does not exceed the threshold, then the signal is interpreted as a process problem and appropriate action is recommended. The benefit of such a scheme is clear – occurrence of unwarranted work stoppage is reduced.

Davis and Adams restrict their analysis to the univariate case, but it is likely that many processes suffering from contamination issues are not characterized by a single quality characteristic, but by several related quality characteristics. A common tool for monitoring several quality characteristics simultaneously is the Hotelling $T^2$ control chart. If the $T^2$ chart is used to monitor a process that has contaminated data, and the chart signals, the analyst must determine if the process is out of control, or if a contaminated sample has caused the chart to signal. We propose
an extension of the diagnostic statistic technique for use with multivariate process monitoring via the Hotelling $T^2$ chart.

**THE $T^2$ CONTROL CHART**

Control charting procedures involve two phases of analysis. A multivariate process is characterized by a mean vector $\mu$ and covariance matrix $\Sigma$ which describes the quality characteristics and their inter-relations. During phase I analysis, an in-control set of data is identified and used to estimate process parameters. The mean vector is estimated by the vector of sample means $\bar{x}$ and the covariance matrix is estimated by the sample covariance matrix $S$.

During phase II analysis, the process is periodically sampled and monitored by plotting one or more statistics on control charts. Unlike univariate control charting procedures, the upper limit of the $T^2$ chart is determined independently of the phase I parameter estimates. The upper control limit of the $T^2$ chart is a multiple of a critical value of an $F$ distribution depending on $p =$ number of quality characteristics, $n =$ subgroup sample size, $m =$ number of phase I samples, and the desired $\alpha$. The phase II upper control limit is given by the following expression:

$$UCL = \frac{p(m + 1)(n - 1)}{mn - m - p + 1} F_{p,n-p-1,m-p+1}$$

Typically, there is no lower control limit in a $T^2$ chart.

The statistic that is plotted is often called Hotelling’s $T^2$ statistic:

$$T^2 = n(\bar{x} - \bar{x})' S^{-1} (\bar{x} - \bar{x})$$

This statistic is essentially the Mahalanobis distance between the mean vector of the sample and the in-control mean vector. Values of $T^2$ that fall above the $UCL$ indicate potentially out-of-control or special cause variation and warrant further investigation. Comprehensive treatment of $T^2$ control charting is given by [2].

**An Example**

Consider monitoring quality of a product with $p = 3$ inter-related quality characteristics. Phase I analysis has been successfully completed and results in the following in-control process parameter estimates:

$$\bar{x} = \begin{bmatrix} 3.034 \\ 3.556 \\ 2.788 \end{bmatrix}$$

$$S = \begin{bmatrix} 1.521 & 1.131 & 1.170 \\ 1.131 & 1.562 & 1.180 \\ 1.170 & 1.180 & 1.315 \end{bmatrix}$$
The phase I estimates are used to establish the formula for the $T^2$ statistics charted during phase II monitoring of future production data. The $UCL$ for the chart is calculated as detailed in the previous section and results in a value of 12.04.

Subsamples of size $n = 10$ are periodically collected from the production process and the $T^2$ statistics are calculated and plotted in Figure 1. The chart signals at sample 37. This is the result of a shift in the mean vector characterizing the process. Samples 38-40 continue to signal.

![Phase II T-sq Chart](image)

Figure 1. $T^2$ chart for the $p = 3$ quality characteristics and subsamples of size $n = 10$. This chart shows signals at samples 37-40.

**The Challenge of Contamination**

We will define “contamination” as a sample of size $n$ containing 1 outlying observation and $n – 1$ typical observations. It is well-known that the sample mean vector and covariance matrix are not resistant to the effects of even a single outlying value in the data. Consequently, contaminated samples can cause a $T^2$ control chart to signal when, in reality, the process is still “in-control” and there is no assignable cause. The average run length ($ARL$) of the $T^2$ control charting scheme would be reduced in this scenario.
Now, consider a process that occasionally produces a contaminated sample as described above. Suppose contamination occurred in sample #10. The manager monitoring the control chart may call for work stoppage after observing the signal from sample #10, but in reality, the process is still in-control and there is no assignable cause. The signal is caused by 1 outlying observation within sample #10. Figure 2 displays the $T^2$ chart for the first 10 samples and 25 additional samples after the signal.

![Phase II T-sq Chart](image)

Figure 2. Sample #10 is contaminated with one outlying observation, but the process is actually in-control, as demonstrated by in-control statistics charted for samples 11-35.

**THE DIAGNOSTIC STATISTIC TECHNIQUE**

We propose a secondary diagnostic statistic (DS) be calculated after the $T^2$ chart signals. The purpose of the DS is to distinguish between signals caused by real changes in the process mean vector and signals caused by a single outlying value within the sample. If the $T^2$ chart signals, the DS is calculated and compared to a decision value. If the DS exceeds the decision value, the signal is deemed to be caused by contaminated data and the process is allowed to continue. If the DS does not exceed the decision value, then the signal is judged to represent a real process change and appropriate action should be initiated.
Diagnostic Statistic Proposals

The value of a DS should reflect the presence or absence of an outlier in the sample under consideration. Contaminated samples should result in large values of the DS. In this paper, we propose two possible diagnostic statistics for use in conjunction with the $T^2$ chart.

Proposal #1: Calculate the mean vector and covariance matrix of the sample data and use these values to calculate the Mahalanobis distance ($MD$) of each observation in the sample to the mean vector of the sample. Choose the maximum of these distances.

$$DS1 = \max(MD_1, MD_2, ..., MD_n)$$

If the maximum value of DS exceeds a designated decision value, conclude that the signal is caused by a data contamination problem.

Proposal #2: Use a “leave-one-out” approach to calculate $n$ sets of Mahalanobis distances (similar to deleted residuals in regression analysis). For samples of size $n$, calculate $n$ sets of “deleted Mahalanobis distances” where the $i$th set of $MD$s is calculated by excluding the $i$th observation from the mean vector and covariance matrix calculation. If the $i$th observation is contaminated, it should stand out as the largest value in the $i$th set of $MD$s. Let $MD_{(i),j}$ represent the Mahalanobis distance of observation $j$ when observation $i$ is the observation left out. Choose the maximum of these $n^2$ distances.

$$DS2 = \max(MD_{(1),1}, MD_{(1),2}, ..., MD_{(n),n})$$

Note that proposal #2 requires subsamples of size at least $n \geq p + 2$. Larger sample sizes are required because $p + 1$ data points uniquely determines an ellipsoid such that these $p + 1$ data points are all exactly the same distance from the mean vector [3]. If these $p + 1$ data points were situated in such a way as to form an elongated, narrow ellipsoid, then the data point that is “left out” could artificially appear as if were an outlier.

Choosing the Decision Value

Since the DS will only be calculated in the event of a signal, the DS technique should use a decision value from the conditional distribution of the DS given a $T^2$ chart signal. We suggest using simulation to calculate an appropriate decision value for use in any given control charting scheme and choice of diagnostic statistic.

We will present a small set of simulations to illustrate the methodology. The following simulations generate “clean” multivariate normal data and run until 10,000 signals are observed. Diagnostic statistic values are recorded for each signal and the decision value is chosen as the $(1 - \alpha)$ percentile of the distribution of DS values. The objective is find a decision value ($dv$) such that $P(DS > dv|T^2 signal) = 1-\alpha$. We restrict this set of examples to $p = 3$, $n = 8$ or 10, and $\alpha$ values of .05, .01, and .005.
<table>
<thead>
<tr>
<th>$\alpha = .05$</th>
<th>$\alpha = .01$</th>
<th>$\alpha = .005$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 8$</td>
<td>5.770</td>
<td>5.964</td>
</tr>
<tr>
<td>$n = 10$</td>
<td>7.016</td>
<td>7.484</td>
</tr>
</tbody>
</table>

Table 1. Decision Value Simulation Results for DS1.

<table>
<thead>
<tr>
<th>$\alpha = .05$</th>
<th>$\alpha = .01$</th>
<th>$\alpha = .005$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 8$</td>
<td>114.6</td>
<td>268.4</td>
</tr>
<tr>
<td>$n = 10$</td>
<td>54.0</td>
<td>99.6</td>
</tr>
</tbody>
</table>

Table 2. Decision Value Simulation Results for DS2.

The practitioner should generate a set of tables reflecting the characteristics of their process data and $T^2$ charting scheme. Computer code for running these simulations in the R statistical computing environment is available from the author upon request.

**$T^2$-Diagnostic Statistic Scheme Example**

Again, consider monitoring quality of a product with $p = 3$ inter-related quality characteristics. Phase I analysis has been completed, in-control parameter estimates are established, and $UCL$ for phase II process monitoring has been calculated. A simulation has been conducted and the decision value for the $DS$ for $p = 3$, $n = 10$, and $\alpha = .01$ has been determined (see Tables 1 and 2 above). Analysts collect subsamples of size $n = 10$ from the production process and the $T^2$ statistics are calculated and plotted in Figure 3.

![Phase II T-sq Chart](image)

Figure 3. Sample #10 is contaminated with one outlying observation. The signals at samples 36-40 are due to a shift in the process mean vector.
The chart signals at sample 10. Both DS1 and DS2 are calculated and both diagnostic statistics exceed the decision values. The analyst monitoring the process can conclude that the signal for sample #10 is due to data contamination, not a shift in the process mean vector and the process is allowed to continue running. The chart signals again at samples 36-40. DS1 and DS2 are calculated for these samples, but the values fall below the decision values, indicating a shift in the process mean is responsible for the signals.

Figure 4 displays the DS1 calculations for all 40 samples and Figure 5 displays the values of DS2 for all 40 samples. (In practice, the DS will only be calculated for those samples that signal, but they are shown for all samples here for illustrative purposes.)

![DS1 Values](image)

Figure 4. Values of DS1 for all 40 samples. Only sample #10 exceeds the decision value.
Figure 5. Values of DS2 for all 40 samples. Only sample #10 exceeds the decision value.

For the contaminated sample, sample #10, the $T^2$ chart signals and the diagnostic statistics exceed the decision values leading the analyst to conclude that the signal is due to data contamination. For samples 36-40, the $T^2$ chart signals, but the diagnostic statistics do no exceed the decision values – the analyst should conclude these signals represent a real shift in the process mean vector and are not due to data contamination – appropriate action should be taken.

**Limitations of the $T^2$-Diagnostic Statistic Scheme**

The proposed scheme is more expensive in terms of data collection than $T^2$ schemes based upon individual’s data. The proposed process monitoring scheme requires collection of subsamples of process data rather than individual’s data. DS1 requires samples of size $n \geq p$. DS2 requires samples of size $n \geq p + 2$.

Another limitation is the number of outliers that the scheme can accommodate. The $T^2$-DS scheme is designed for the specific situation of occasional samples containing a single outlying value. If the measurement system is so problematic that samples are contaminated with multiple outlying values, the proposed scheme will be less useful.

**CONCLUSIONS**

The $T^2$ chart can be augmented by a secondary diagnostic statistic to effectively monitor multivariate process data in the presence of occasional data contamination. If subsamples are used in the $T^2$ control charting scheme, then the DS can help the analyst distinguish between true shifts in the process parameters and signals caused by single outliers.
REFERENCES

