Chapter 2

Multivariate Statistical Process Control

2.1 Multivariate Data

It is assumed that the Phase I historical data set consists of m time ordered vectors that are independent of each other. Each vector is of dimension p, so $\mathbf{a}_i$ is a vector containing p elements for the i^{th} time period. When the process is in-control, each $\mathbf{a}_i$ is assumed to come from the same multivariate normal distribution, that is, $\mathbf{a}_i \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here $\boldsymbol{\mu}$ is the population mean vector that determines a point in p-dimensional space that represents the location and $\boldsymbol{\Sigma}$ is a p by p positive definite variance-covariance matrix that determines the dispersion, which is also referred to as scatter or shape. The two major types of instability that we consider in this research are outliers and step changes. Methods that work well for these two types of instability will often work well for other types of instability that are more difficult to study.

Outliers

Outliers in multivariate data are more difficult to detect than outliers in univariate data.

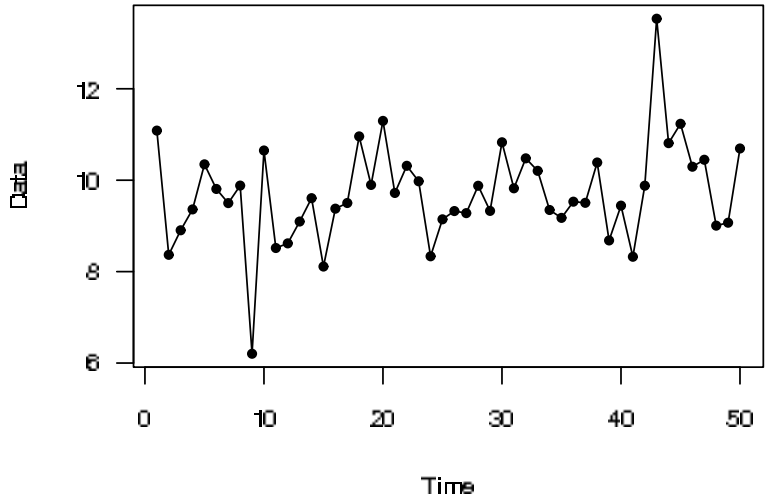
One reason for this is because simple graphical methods that can be used to detect univariate outliers are often not available in higher dimensions. Another reason is because there are many more ways that the multivariate data can come from an out-of-control process. For example, there could be outliers due to changes of location in random directions for each outlier, a cluster of outliers due to a location shift in a particular direction, multiple clusters of outliers in different directions, points with the same location as the good data but with more variability, or the outliers can be due to a shift in some of the elements of the location vector but not all of them. The term "masking effect" has been coined to describe the situation where multiple outliers are present and inflate the variance-covariance estimates in such a way that they mask each other and escape detection. See Rocke and Woodruff (1993) for a discussion of various types of outliers.

Rocke and Woodruff (1996) stated that the most difficult type of multivariate outliers to detect are those that have the same variance-covariance matrix as the good data. These difficult-to-detect outliers are referred to as "shift outliers" because the center of the outlying points has been shifted by some distance from the center of the good data. The categorization of shift outliers includes individual points as well as clusters of points. If shift outliers can be detected by the particular estimation method, then the method will likely work well for other kinds of outliers, hence the focus on shift outliers here. Figure 2.1 gives an example of time-ordered quality control statistics where outliers are present. The plotted statistics for each time period can represent a univariate measure or a multivariate measure that combines several components into a single statistic.

Step Changes

A step change occurs when something has happened in the process so that the mean vector, $\boldsymbol{\mu}$, has changed to some new vector, $\boldsymbol{\mu_1}$. We assume that $\boldsymbol{\Sigma}$ remains the same

Figure 2.1: A sample of quality control data with outliers.

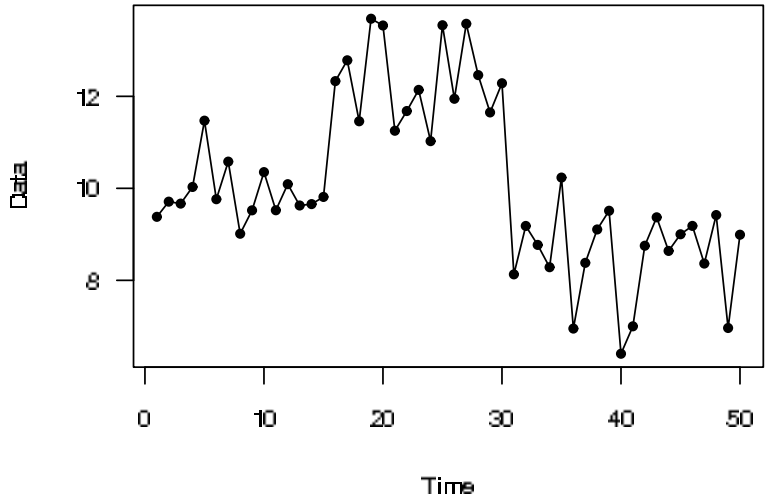


when a step change occurs and that the step change occurs at some time between two successive observations. The step change remains for the rest of the Phase I data. While these assumptions may not always hold in practice, they are used here to simplify the properties of the control chart and makes for easier comparison of analysis methods. Charts that perform well for step changes often perform well for other types of changes, such as temporary step changes or for an increasing trend. Figure 2.2 gives an example of quality control data with two distinct step changes.

2.2 T^2 Statistic

Identification of outliers can be done with the T^2 statistic with a single control limit which is widely used for multivariate data analysis. A comprehensive review of the T^2 statistic, its properties and alternative forms can be found in Mason and Young (2002). The general

Figure 2.2: A sample of quality control data with step changes.



form of the statistic is

$$T_i^2 = (\mathbf{a}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{a}_i - \boldsymbol{\mu}) \text{ for } i = 1, 2, \ldots, m, \tag{2.1}$$

which has a χ_p^2 distribution when $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known.

Because $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are usually not known, they are replaced with appropriate estimators. The classical estimators are the sample mean vector and sample variance-covariance matrix given by,

$$\overline{\mathbf{a}} = \frac{\sum_{i=1}^{m} \mathbf{a}_i}{m}, \tag{2.2}$$

and

$$\mathbf{S}_1 = \frac{\sum_{i=1}^{m} (\mathbf{a}_i - \overline{\mathbf{a}}) (\mathbf{a}_i - \overline{\mathbf{a}})'}{m-1}. \tag{2.3}$$

It is important to realize that once an estimator is used in place of parameters that the T^2 statistics are no longer independent of each other in Phase I because they the data are often not independent of the estimator based on the data. For Phase II, the future observations are independent of the estimators but probability calculations involving the

plotted statistics cannot assume independence because they are based on the same Phase I control limit. Discussion of this idea can be found in Jensen et al. (2006). The correlation of the T^2 statistics calculated from Phase I data was given as $1/(m-1)$ in Mason and Young (2002, p. 25).

A T_i^2 statistic based on these classical estimators in (2.2) and (2.3) is denoted by $T_{1,i}^2$ and values are given by

$$T_{1,i}^2 = (\mathbf{a}_i - \overline{\mathbf{a}})' \, \mathbf{S}_1^{-1} \, (\mathbf{a}_i - \overline{\mathbf{a}}) \text{ for } i = 1, 2, \ldots, m. \tag{2.4}$$

This statistic is equivalent to the squared Mahalanobis distances and has been shown to be effective in detecting a single moderately-sized multivariate outlier as shown in Figure 1 of Vargas (2003). In addition, it can be shown that the distribution of $T_{1,i}^2$ is proportional to a beta distribution (Chou, Mason, and Young, 1999; Atkinson, Riani, and Cerioli, 2004), that is,

$$T_{1,i}^2 \frac{m}{(m-1)^2} \sim Beta\left(\frac{p}{2}, \frac{m-p-1}{2}\right) \text{ for } i = 1, 2, \ldots, m. \tag{2.5}$$

This known distributional result makes it easy to calculate a control limit for $T_{1,i}^2$, assuming that the sample size m is large enough so that the correlation of the T^2 statistics has little effect. Justification for why the beta distribution, a bounded distribution, is found in the fact that $\sum_{i=1}^{m} T_{1,i}^2 = p(m-1)$ (Atkinson, Riani, and Cerioli, 2004. p. 44).

Note that this distributional result in (2.5) holds as long as the data are independent and identically distributed (i.i.d.) with a multivariate normal distribution. The matrix $\mathbf{\Sigma}$ can be any positive definite variance-covariance matrix. Thus when the observations within a vector are correlated with each other, the $T_{1,i}^2$ statistics will still be proportional to a beta distribution.

However, as shown by Sullivan and Woodall (1996), use of the estimator in (2.3) is not effective in detecting sustained step changes in the mean vector, nor is it effective in detecting

multiple outliers (Vargas, 2003). While $T^2_{1,i}$ has been shown to be effective in detecting a single moderately-sized multivariate outlier as shown in Figure 1 of Vargas (2003), a single arbitrarily large outlier or step change can render the $T^2_{1,i}$ statistic useless. Chou, Mason and Young (1999) recommended the $T^2_{1,i}$ statistic to detect outliers but it is not clear the number of outliers that they used in their simulation studies nor is it clear that their simulation results are correct. Based on our studies shown in Chapter 5, we concur with the conclusions of Sullivan and Woodall (1996) and Vargas (2003) and do not recommend the use of the $T^2_{1,i}$ statistic for Phase I analysis when outliers or step changes may be present.

2.3 Alternative T^2 Statistics

An alternative is to base the T^2_i statistics on the sample mean vector and the variance-covariance matrix estimated using the successive differences between vectors, denoted by $T^2_{2,i}$ (Holmes and Mergen, 1993). If $\mathbf{v}_i = \mathbf{a}_{i+1} - \mathbf{a}_i$ is the vector of the i^{th} successive difference, then an unbiased estimator of the variance-covariance matrix is

$$\mathbf{S}_2 = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \mathbf{v}_i \mathbf{v}_i^{'}. \tag{2.6}$$

This statistic is analogous to the use of the moving range to construct an univariate Shewhart Individuals chart. Sullivan and Woodall (1996) showed that $T^2_{2,i}$ is effective in detecting sustained step changes in the process that occur in Phase I data. While the distribution of $T^2_{2,i,MIX}$ does not have a simple closed form, its asymptotic distribution is χ^2_p. A discussion of the various approximate distributions and the preferred χ^2_p approximation for large samples is given in Williams et al. (2006b). The sample sizes that we use here are large enough to justify use of the χ^2_p approximation to obtain the control limit. However, like $T^2_{1,i,MIX}$, $T^2_{2,i,MIX}$will not be effective in detecting multiple multivariate outliers (Vargas, 2003).

Robust alternatives of the T^2 statistics considered here are based on either the minimum volume ellipsoid (MVE) estimators or the minimum covariance determinant (MCD) estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. These will be denoted by $T^2_{mve,i}$ and $T^2_{mcd,i}$ respectively, and defined as

$$T^2_{mve,i} = (\mathbf{x}_i - \mathbf{x}_{mve})' \, \mathbf{S}^{-1}_{mve} \, (\mathbf{x}_i - \mathbf{x}_{mve}) \, i = 1, 2, \ldots, m, \tag{2.7}$$

$$T^2_{mcd,i} = (\mathbf{x}_i - \mathbf{x}_{mcd})' \, \mathbf{S}^{-1}_{mcd} \, (\mathbf{x}_i - \mathbf{x}_{mcd}) \, i = 1, 2, \ldots, m. \tag{2.8}$$

where $\mathbf{x}_{mve}$ and $\mathbf{x}_{mcd}$ are the corresponding location estimators and $\mathbf{S}_{mve}$ and $\mathbf{S}_{mcd}$ are the corresponding estimators of the variance-covariance matrix. In Chapter 3 we discuss these robust estimators in more detail, explain how they are calculated, and show when it is preferable to use them.

2.4 Other Multivariate Charts

For those familiar with multivariate quality techniques, it is often thought that alternative multivariate control charts, such as the multivariate exponentially weighted moving average (MEWMA) or multivariate CUSUM (MCUSUM) charts will be useful. However, for Phase I applications, it is best to use charts that do not use prior information as the MEWMA and MCUSUM do. They are designed to detect small or gradual changes in the mean vector and work well for Phase II application. They do not work as well as a T^2 based chart for larger outliers and step changes in Phase I applications. In addition, if a signal is present on a MEWMA or MCUSUM chart, it is not clear which point(s) is different from the others and no guidance is given on how to clean the Phase I dataset in preparation for Phase II applications. This issue was noted for univariate charts such as the exponentially weighted moving average (EWMA) by Kim, Mahmoud, and Woodall (2003).