MULTIVARIATE STATISTICAL PROCESS CONTROL FOR CORRELATION MATRICES

by

Mark F. Sindelar

B. S. in Electrical Engineering, University of Akron, 1995

M. B. A., John Carroll University, 1999

M. S. in Industrial Engineering, University of Pittsburgh, 2001

Submitted to the Graduate Faculty of the

School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

SCHOOL OF ENGINEERING

This dissertation was presented

by

Mark F. Sindelar

It was defended on

March 1, 2007

and approved by

Mainak Mazumdar, Professor Emeritus, Industrial Engineering

Larry J. Shuman, Professor & Associate Dean for Academic Affairs, School of Engineering

Pandu R. Tadikamalla, Professor, Business Administration

Harvey Wolfe, Professor Emeritus, Industrial Engineering

Dissertation Director: Mary E. Besterfield-Sacre, Associate Professor, Industrial Engineering

MULTIVARIATE STATISTICAL PROCESS CONTROL

FOR CORRELATION MATRICES

Mark F. Sindelar, Ph.D.

University of Pittsburgh, 2007

Measures of dispersion in the form of covariance control charts are the multivariate analog to the univariate *R*-chart, and are used in conjunction with multivariate location charts such as the Hotelling T^2 chart, much as the *R*-chart is the companion to the univariate *X*-bar chart. Significantly more research has been directed towards location measures, but three multivariate statistics ($|\boldsymbol{S}|$, W_i, and G) have been developed to measure dispersion. This research explores the correlation component of the covariance statistics and demonstrates that, in many cases, the contribution of the correlation component is less significant than originally believed, but also offers suggestions for how to implement a correlation control chart when this is the variable of primary interest.

This research mathematically analyzes the potential use of the three covariance statistics ($|\boldsymbol{S}|$, W_i, and G), modified for the special case of correlation. A simulation study is then performed to characterize the behavior of the two modified statistics that are found to be feasible. Parameters varied include the sample size (n), number of quality characteristics (p), the variance, and the number of correlation matrix entries that are perturbed. The performance and utility of the front-running correlation (modified W_i) statistic is then examined by comparison to similarly classed statistics and by trials with real and simulated data sets, respectively. Recommendations for the development of correlation control charts are presented along with a description of the

types of process to which they apply. An outgrowth of the research is the understanding that the correlation component often does not contribute as much as the scale factor component of the dispersion measure in many cases.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation is dedicated to Wilma Powell, for her encouragement, support, understanding and friendship.

My father, Kenneth F. Sindelar, was the first engineer in my life and the one that deserves credit for introducing me at an early age to the basic concepts of Engineering which would become the foundation of my schooling and profession; my mother, Mary Ann V. Sindelar deserves thanks for her patience, reassurance, and encouragement through my many decades of education.

United States Steel Corporation, and particularly my Engineering Manager, Raymond W. Boronyak, have made numerous accommodations to assist my pursuit of this degree. This cooperation has been very much appreciated.

I am indebted to my advisor and dissertation committee chair, Mary E. Besterfield-Sacre and am thankful for the contributions of committee members Mainak Mazumdar, Larry J. Shuman, Pandu R. Tadikamalla, and Harvey Wolfe.

I am also thankful for the assistance of Murat Caner Testik, Associate Professor and Department Chair, Industrial Engineering, of Hacettepe University, who reviewed the derivations of Chapter 3 and offered helpful suggestions.

Finally, I thank Robert S. Parker, Associate Professor of Chemical Engineering at the University of Pittsburgh, for use of the simulator software of Section 7.3, and the many fine people in the University Library System for their assistance obtaining resources.

1.0 INTRODUCTION

Multivariate quality control charts expand the options for process monitoring beyond the traditional univariate case by allowing for the simultaneous monitoring of several variables. In the multivariate setting, many developments have provided improved charts for monitoring location (*i.e.* the mean vector) while relatively little research has addressed measures of dispersion (Alt, 1985; Golnabi & Houshmand, 1996). However, it has been recognized that dispersion may have a significant impact on the location vector (Hayter & Tsui, 1994; Houshmand *et al.*, 1997).

Thus, while three statistics to measure dispersion exist ($|\boldsymbol{S}|$, W_i, and G), a comparison has not been made among them to determine their effectiveness for a variety of process conditions. The sample generalized variance, $|\boldsymbol{S}|$, is the oldest of the statistics, but several authors have cited possible disadvantages. The W_i statistic was introduced by Alt (1985) and appears to overcome many of the drawbacks reported with using the $|\boldsymbol{S}|$ chart. More recently, the G chart has been introduced by Levinson *et al.* (2002) and incorporates mean square successive differences in an attempt to increase sensitivity of the dispersion measure. Each of these statistics can be plotted in a traditional quality control manner against time or lot number to monitor a process as shown in Figure 1 where "CL" (center line) is the mean value and "UCL" and "LCL" are the upper and lower control limits, respectively. If the plotted point exceeds the control limits (as shown by the last point) the statistic is out-of-control, indicating a possible change in the process.

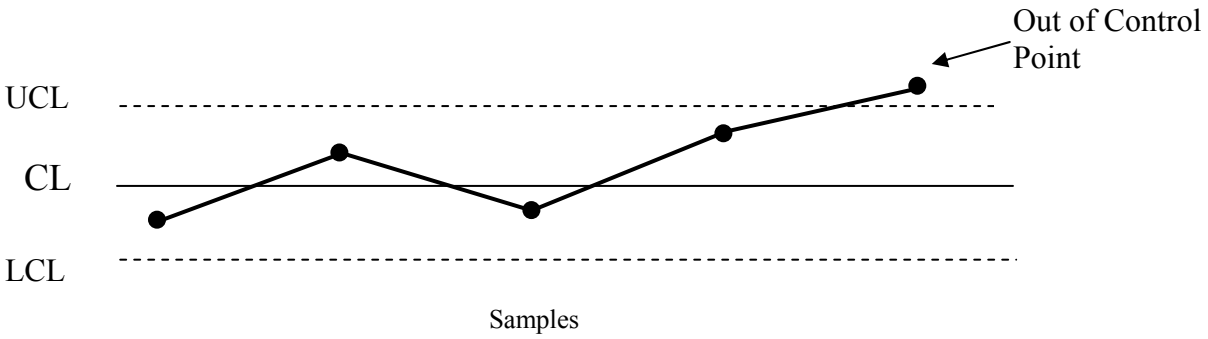

Figure 1 Control Chart

Multivariate dispersion charts are usually intended to supplement location charts, much as the univariate *R*-chart is used as a supplement to the *X*-bar chart. Despite this orientation, several authors have indicated that dispersion, in the form of the variance-covariance matrix, is often a component of location and, therefore, affects changes in the mean vector (Lowry and Montgomery, 1995; Golnabi and Houshmand, 1996). Understanding the behavior of covariance independent of the mean vector in a statistical quality control setting guides the direction of this research.

As an avenue to begin an investigation, consider instead the correlation matrix, a standardized variance-covariance matrix, for the same data set. The diagonal elements, each representing the correlation of a number with itself, are all equal to unity and provide an upper boundary condition. This bounded nature of the correlation matrix provides an alternative and initial approach to the comparison of dispersion statistics. Furthermore, this approach also separates the covariance matrix into two components: the correlation matrix and a scale factor. The contribution of the correlation component to the dispersion measure can, thus, also be evaluated in this manner and this is the focus of this dissertation.

The generation and use of correlation control charts rather than covariance charts may have some advantages in application as well as contribute to understanding dispersion in

Multivariate Statistical Process Control (MSPC). Similar to the variance-covariance case, changes in the correlation matrix could be used to help explain changes in the location vector that appear on a location chart such as the Hotelling T^2 chart (1947). A chart that monitors for changes in the correlation matrix for p quality characteristics could also be used to monitor for changes in the dispersion, specifically the correlation, matrix separate from the location vector.

Additionally, if the data for the in-control correlation matrix originates from a process where no changes are expected, then changes in the sampled correlation matrix may indicate that the underlying assumptions have changed and a new analysis for the current data would likely reveal the discrepancies. This is particularly applicable to process data such as that found in the chemicals industry where certain relationships are assumed to remain independent when the process is stable.

1.1 OBJECTIVES OF THE RESEARCH

This research investigates the modification of known statistics for covariance ($|\boldsymbol{S}|$, W_i, and G) to the special case of correlation to: (1) compare and evaluate the three statistics for a number of parameters; (2) provide guidance for the application of said statistics to correlation control charts; (3) investigate the application of correlation control charts; and (4) make assessments of the role correlation plays in dispersion.

1.2 RATIONALE FOR THE DEVELOPMENT OF CORRELATION CONTROL CHARTS

This research focuses on evaluation of the dispersion statistics used for development of multivariate control charts where the unique nature of correlation matrices would make the correlation control chart itself a tool in certain circumstances where the measures of linear association are of interest to an organization.

For example, in the chemicals industry a distillation column is expected to produce outputs which are correlated with one another because of their stoiciometric relationship. A change in the feedstock composition could change the correlation between outputs if this balance is upset. All the outputs of the column may be saleable but, if their relation changes due to a change in the input feedstock, then the column operator should be made aware.

This scenario also illustrates an important observation for interpretation of correlation control charts. Namely, it is the proposition that a shift in the correlation matrix does not necessarily denote an out-of-control condition in the traditional sense of the term. The shift may be attributable to a change in requirements, as opposed to requirements not being satisfied. Thus, the study of correlation, the standardized version of the variance-covariance matrix, provides additional value in its conceptual form.

1.3 ORGANIZATION OF THE WORK

To begin an exploration of correlation as a measure of dispersion, it is illustrative to start with the first MSPC chart and statistic. In Chapter 2, the Hotelling T^2 chart, which considered the

location vector, is introduced. Since its inception in 1947, additional research has modified and improved upon Hotelling's approach to the monitoring of the location vector. Charts for dispersion then followed, albeit less in number. The Literature Review presented in Chapter 2 places particular emphasis on the dispersion statistics for covariance ($|\boldsymbol{S}|$, W_i, and G) that are considered in modified form in this research.

The evaluation of the three statistics in Chapter 3 begins with a mathematical analysis to determine the applicability of the covariance statistics modified to the correlation special case and investigates the associated control limits. From these analyses, the $|\boldsymbol{S}|$ and W_i statistics in modified form for correlation emerge as feasible possibilities for the construction of control charts. The bounded nature of the correlation matrix, while providing a framework for the analysis, also raises some concerns about the sensitivity of the control limits in certain situations.

Chapter 4 starts with a simulation study to evaluate the correlation case for various levels of quality characteristics (p) and sample size (n) for two common matrices. A terminating sequential simulation provides data with the performance measure being the Average Run Length, or *ARL*. The results of the simulation indicate that only a modified form of the W_i statistic (noted as the W_R statistic) is possible, under specified conditions, for the development of correlation control charts. An empirical equation developed from the simulation results is then used to characterize the behavior of this statistic.

A comparison with similarly developed MSPC statistics is presented in Chapter 5, along with a discussion of the effect that each parameter has on the statistic. Existing data sets are then employed in Chapter 6 to demonstrate possible applications with results that vary from encouraging to discouraging. The insight gained from attempting to control chart these data sets

allows for a summarization of implementation issues and used to make conclusions about the utility of this branch of MSPC.

In Chapter 7, recommendations for associated future work and the direction thereof are then presented. A summary of the work is also provided, showing how the W_R statistic for correlation, developed from the W_i statistic for covariance, is really the only feasible alternative for control charting. While this statistic behaves similar to other statistics of its type, it also similarly has a number of issues with regard to practical implementation. However, the upshot of the conclusion is that the correlation component of the variance-covariance matrix is often of less influence than the scale factor component.

2.0 LITERATURE REVIEW & BACKGROUND

The literature review presented is partitioned into four areas. First an overview of multivariate statistical process control and specifically the most commonly used multivariate chart is discussed in Section 2.1. Next, in Section 2.2, a summary of the research associated with multivariate control charts is presented. In section 2.3, the three statistics of dispersion specific to this research are discussed in depth. Other dispersion measures are briefly discussed in Section 2.4. Section 2.5 provides the background and rationale for the use of the correlation component of the covariance matrix.

2.1 MULTIVARIATE STATISTICAL PROCESS CONTROL

Following the original univariate mean and range charts of Shewhart are a variety of multivariate control charts. Perhaps the best known is the Hotelling T^2 chart, the more common version of the χ^2 chart in which the covariance matrix and mean vector are not known and are estimated from the current sample (Montgomery, 1997). A single statistic is generated by the equation

$$T^2 = n(\overline{X} - \overline{\overline{x}})' S^{-1} (\overline{X} - \overline{\overline{x}}) \qquad (2\text{-}1),$$

where n is the sample size, $\bar{\bar{x}}$ is the estimate of the true mean vector, μ, and $\boldsymbol{S}$ is the estimate of the true covariance matrix, $\boldsymbol{\Sigma}$. This statistic is plotted as shown in Figure 1, against control limits of the form

$$\begin{cases} UCL = \dfrac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \\ LCL = 0 \end{cases} \qquad (2\text{-}2),$$

where p represents the number of quality characteristics, n is the sample size for each group, m represents the number of sample groups and the F-statistic comes from the distribution with the number of degrees of freedom as specified.

2.2 A SUMMARY OF LITERATURE ASSOCIATED WITH THE HOTELLING T^2 CONTROL CHART AND ITS DEVELOPMENTS

The concept of developing a correlation control chart, with an objective of characterizing covariance behavior, evolves naturally in the progression of research into multivariate charts that originated with the Hotelling T^2 chart. Figure 2 provides an overview of these research directions that are reviewed herein. Most research since the late 1950s involving the T^2 chart has concentrated on monitoring changes in the location or mean vector, as depicted on the left side of Figure 2. The developments for location (mean) can be subdivided into two general areas: determination and tolerance. Determination denotes research concerned with identifying univariate components responsible for generating an out-of-control point on the T^2 chart. Tolerance applies to research seeking the preferred method for assigning control limits to the T^2

chart. Similarly, as shown on the right side of Figure 2, there have also been developments investigating changes in the dispersion (variance), including two hybrid methods that also incorporate information related to the location vector. Figure 2 is explained in the Literature Review with the emphasis placed primarily on the dispersion measures.

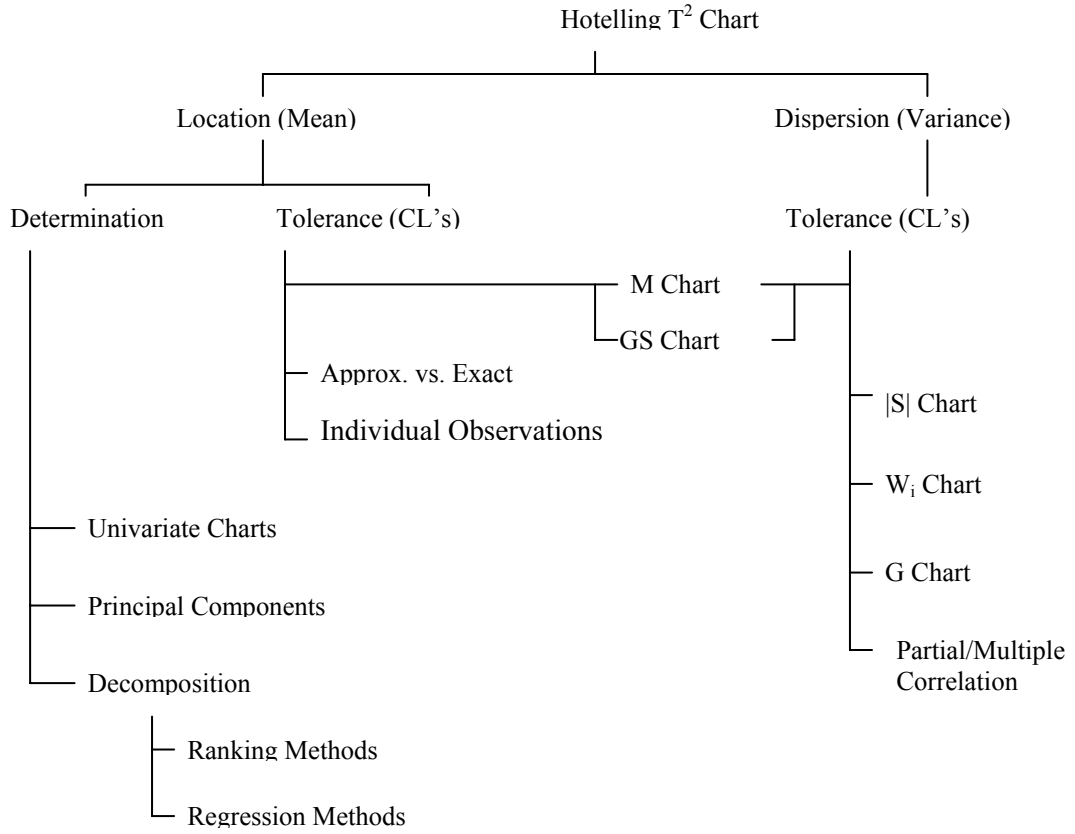

Figure 2 Research Directions Related to the Hotelling T^2 Chart

2.2.1 Research on Control Limits for Location

Research on the T^2 chart mentions several alternatives for setting control limits. Typically, Phase I limits are used to estimate process parameters from the sample when exact values are not known, and then Phase II limits are established for the process in control (Alt, 1985, Sullivan and Woodall, 1995). However, if the process mean and variance are known *a priori* it is generally

accepted that Phase II limits can be applied. In general, Phase II control limits are established with an upper limit based on the chi-squared distribution with p degrees of freedom and the lower control limit at zero.

Much of the research investigates the differences between using asymptotic approximations and exact values in the setting of Phase I control limits and the asymptotic estimates are found to be acceptable for most cases (Alt, 1985). Tracy *et al.* (1996) have made a specific contribution in the area of MSPC addressing the problem of single observations plotted on the T^2 chart. That is, they address the situation where the sample size, n, equals one. Sullivan and Woodall (1996) review various approaches for single sample control charts, including the work of Tracy *et al.* related to the T^2 chart.

2.2.2 Research on the Explanation of Out-of-Control Points

A simple method for determination of out-of-control points is plotting a univariate control chart for each variable used in calculation of the T^2 statistic. The upper and lower control limits are calculated using a Bonferroni approach so that the tolerance on each univariate statistic is α/p, where p is the number of variables and α is the overall simultaneous tolerance requirement.

The earliest alternative approach to determining the out-of-control components involved decomposition by principal components analysis (PCA) and stems from the work of Jackson (1959, 1980, 1985). The chemical process industry readily uses this approach but more recent techniques that more accurately define out-of-control variables can be found in the literature. As noted by Hayter and Tsui (1994), the variables resulting from PCA are often quantities such as "(1/4) weight + (3/4) length" that do not have an analog in the process.

Two interrelated methods have recently been used in place of principal components analysis to determine those variables which cause an out-of-control signal on the T^2 chart. Doganaksoy *et al.* (1991) introduced a method that ranked the components of the observation vector based on a statistic compared to a *t*-distribution. Similarly, Hawkins (1991, 1993) has done extensive work to improve the capabilities of multivariate control charts by making regression adjustments to individual variables. Hawkins (1993) and Wade and Woodall (1993) have applied these regression adjustments to the T^2 chart to identify influential components. In 1995, Mason *et al.* demonstrated a comprehensive decomposition procedure by which the regression and ranking techniques were both shown to be subsets.

2.2.3 Research on Control Limits for Dispersion

Referring back to Figure 2, the other factor influencing T^2 charts concerns dispersion. It has been recognized that much work remains in this area, including a 1995 review of multivariate control charts by Lowry and Montgomery and an article by Golnabi and Houshmand (1996). Of particular note in the existing literature is the orientation, although it is not explicitly stated, that changes in the variance-covariance matrix (or correlation matrix at its most basic state) are considered nuisances that must be addressed in order to ensure that the T^2 chart is properly monitoring the process location. The *M* chart (Hayter and Tsui, 1994) and *GS* chart (Houshmand *et al.,* 1997) combine dispersion with location to address this issue. Research that directly attacks the dispersion question has resulted in the production of three statistics for control charting purposes: the $|\boldsymbol{S}|$, the W_i, and the G. In addition, the methods of partial and multiple correlation have been introduced. The latter methods, while providing the ability to indicate in- and out-of-control conditions, are algorithmic applications (Golnabi and Houshmand, 1996).

These methods were not designed with the intent of adaptation to control charts and do not produce a single, chartable statistic. Thus, they are only briefly summarized. The following section elaborates on the development of the three statistics ($|\boldsymbol{S}|$, W_i, and G) of interest to this research.

2.3 DISPERSION CONTROL CHARTS

As previously noted, three dispersion statistics ($|\boldsymbol{S}|$, W_i, and G) and associated control charts have been developed. Each is described below in the chronological order of their introduction—this facilitates understanding the motivation for the more recent developments.

2.3.1 The $|\boldsymbol{S}|$ Chart

The sample generalized variance, $|\boldsymbol{S}|$, is one of the most widely used measures of process dispersion (Alt, 1985). The foundation of the $|\boldsymbol{S}|$ chart is based on the assumption that the determinant of the covariance matrix is the multivariate analog to covariance (Wilks, 1932).

One attraction of the $|\boldsymbol{S}|$ chart is ease of calculation and the resultant scalar $|\boldsymbol{S}|$ that is plotted against the control limits. Alt (1985) notes that the ". . . $|S|$-control chart can be constructed using only the first two moments of $|S|$ and the property that most of the probability distribution of $|S|$ is contained in the interval $E(|S|) \pm 3\sqrt{V(|S|)}$" (p. 116).

The control limits are

$$\begin{cases} UCL = |\Sigma_0| \dfrac{\left(\chi^2_{2n-4,\alpha/2}\right)^2}{4(n-1)^2} \\ LCL = |\Sigma_0| \dfrac{\left(\chi^2_{2n-4,1-(\alpha/2)}\right)^2}{4(n-1)^2} \end{cases} \quad (2\text{-}3),$$

where Σ_0 is the in-control covariance matrix, n is the sample size, and α is the specified tolerance. If $n < 6$, the lower control limit is replaced with zero. The control limits in equation 2-3 are more common, although others exist for specific purposes (Golnabi and Houshmand, 1996; Alt, 1985; Montgomery and Wadsworth, 1972).

Despite ease of calculation and intuitive plausibility, the $|\boldsymbol{S}|$ statistic has several potential drawbacks as a relatively simplistic, scalar representation of a complex multivariate structure (Lowry and Montgomery, 1997). Alt (1985) cites an example from Johnson and Wichern using bivariate data resulting in "distinctly different correlations, r = 0.8, 0.9, and -0.8" although the sample covariances have the same generalized variance (p. 116). Similarly, Lowry and Montgomery (1997) note three covariance matrices:

$$S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (2\text{-}4a)$$

$$S_2 = \begin{pmatrix} 2.32 & 0.40 \\ 0.40 & 0.50 \end{pmatrix}, \text{ and} \quad (2\text{-}4b)$$

$$S_3 = \begin{pmatrix} 2.65 & -0.40 \\ -0.40 & 0.50 \end{pmatrix}, \quad (2\text{-}4c)$$

each of which "conveys considerably different information about process variability and the correlation between the two variables" but result in $|\boldsymbol{S}_1| = |\boldsymbol{S}_2| = |\boldsymbol{S}_3| = 1$ (p. 804). Consequently,

Lowry and Montgomery (1997), Alt (1985) and others suggest that the $|\boldsymbol{S}|$ chart not be used in isolation, but rather be used in conjunction with univariate dispersion control charts created from the multivariate data.

2.3.2 The W_i Chart

Some concerns related to use of the sample generalized variance, $|\boldsymbol{S}|$, can be alleviated by employing a direct multivariate extension of the univariate S^2 control chart. The evolution of this control chart started with Alt in the mid-1970s and since 1985 it has been regularly mentioned in the literature as an alternative to the $|\boldsymbol{S}|$ chart.

Assuming that the true covariance matrix, $\boldsymbol{\Sigma}$, is known (or estimated from a large, in-control sample), multiple comparisons are made between the sample covariance matrices from the process and the known covariance matrix. The multiple comparisons are a series of tests of significance of the form H_0: $\sigma^2 = \sigma_0^2$ vs. H_1: $\sigma^2 \neq \sigma_0^2$ (Alt, 1985, p. 116). In this case, the repeated hypothesis tests compare the known and the sample covariance matrices.

The test statistic computed and plotted for each sample, i, is

$$W_i = -pn + pn\ln(n) - n\ln\left(\frac{|A_i|}{|\Sigma|}\right) + tr(\Sigma^{-1}A_i) \tag{2-5}$$

where

$$A_i = (n-1)S_i \tag{2-6}$$

(Lowry and Montgomery, 1997, p. 804). In equations 2-5 and 2-6, p denotes the number of quality characteristics, n denotes the sample size, and $\boldsymbol{S}$ is the sample covariance matrix for the i^{th} sample.

The calculated point, W_i, is compared to an upper control limit (UCL) with a χ^2 distribution with $p(p+1)/2$ degrees of freedom at a significance level of α, as shown in equations 2-7,

$$\begin{cases} UCL = \chi^2_{\left(\frac{p(p+1)}{2},\alpha\right)} \\ LCL = 0 \end{cases} \quad (2\text{-}7).$$

2.3.3 The *G* Chart

The *G* Chart is a recent application of the *G* statistic to multivariate statistical process control for dispersion introduced by Levinson *et al.* (2002). This work stems from attempts to improve the T^2 statistic (Holmes and Mergen, 1993) in view of variation and starts with the q^2 statistic reported by Hald (1952). Hald showed that q^2 is an unbiased estimator of σ^2 given by the equation:

$$q^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2 \quad (2\text{-}8),$$

where X_i and X_{i+1} are consecutive sample values, and n is the sample size. Equation 2-8 represents the Mean Squared Successive Differences (MSSD) estimation of the variances. Hald also proposed a ratio r (not to be confused with the correlation coefficient, r) that is equal to q^2/s^2. He suggested that this ratio could be used as an "hypothesis alternative to the hypothesis of statistical control" that would indicate a gradual change in the population mean, with small

values of r being significant (Hald, p. 358). The distribution of r is approximately normal for $n > 20$; a table of fractiles is presented for values from $n = 4$ to $n = 20$.

Holmes and Mergen (1993) applied the concept of MSSD as an attempt to improve the sensitivity of the T^2 control chart (p. 261). Since q^2 is an unbiased estimator of σ^2, it can be used in place of s^2 to calculate the variance-covariance matrix $\boldsymbol{S}$. Consider two processes, X and Y, and the covariance calculated by the standard method and by the MSSD method. "If the X and Y processes are random and have no cross-correlation other than at zero lag, these two estimates of the universe covariance will be similar. If these conditions are not met, then there will be a difference in the estimators" (p. 622).

Holmes and Mergen (1998) extend the MSSD application as "a test for the equality of the regular covariance matrix and the MSSD covariance matrix to establish whether or not the multivariate process is stable" (p. 505). It employs a G statistic to compare the equality of two population covariance matrices calculated as

$$G = 2.3026(m)(M), \tag{2-9}$$

where m is a constant defined as

$$m = 1 - \left[\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1}\right) - \left(\frac{1}{n_1 + n_2 - 2}\right)\right]\left(\frac{2p^2 + 3p - 1}{6(p+1)}\right) \tag{2-10}$$

and

$$M = (n_1 + n_2 - 2)\log|S| - (n_1 - 1)\log|S_1| - (n_2 - 1)\log|S_2| \tag{2-11}$$

where p represents the number of quality characteristics, n is the sample size, and $\boldsymbol{S}$ is the covariance matrix. Subscripts refer to the individual data for each of two covariance matrices, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ from samples of size n_1 and n_2, respectively, so that $\boldsymbol{S}$ represents a pooled covariance. In this equation the logarithm used is base 10.

The G statistic was proposed by Kramer and Jensen (1969b) and is based on the assumption that the determinant of the covariance matrix is the multivariate analog of the variance (Holmes and Mergen, 1998). It should be noted that this assumption comes from Wilks (1932), and is subject to the same arguments presented in Section 2.3.1.

Levinson *et al.* (2002) reason that a stable process would produce separate estimates of the covariance matrix that are approximately equal, and that the equivalence of these matrices could be tested using the G statistic. Two methods are presented for calculation of the covariance matrix, $\boldsymbol{S}$. The first is the standard calculation, referred to by Levinson *et al.* as the "full data set" method. The second method uses the MSSD (Holmes and Mergen, 1998) and is given as

$$S_{MSSD} = \frac{1}{2(n-1)} \sum_{j=2}^{n} (X_j - X_{j-1})(X_j - X_{j-1})^T \qquad (2\text{-}12),$$

where X_{j-1} and X_j are consecutive samples in a sample of size n, computed for k subgroups (number of samples), and T is the transpose operator. The value of the constant m remains the same as in equation 2-10 above, but the value of M changes to

$$M = (\nu_1 + \nu_2) \ln |S| - \nu_1 \ln |S_1| - \nu_2 \ln |S_2| \qquad (2\text{-}13)$$

where $|\boldsymbol{S}_2|$ is the sample generalized variance for the sample being compared to the in-control sample generalized variance, $|\boldsymbol{S}_1|$, and $\boldsymbol{S}$ is the pooled variance given by

$$S = \frac{(\nu_1 S_1 + \nu_2 S_2)}{\nu_1 + \nu_2} \qquad (2\text{-}14)$$

where

$$\nu_1 = k(n_2 - 1) \qquad (2\text{-}15a)$$

and

$$\nu_2 = (n_2 - 1) \tag{2-15b}$$

where ν_1 and ν_2 are the degrees of freedom for the initial (control) sample, and subsequent samples, respectively, with n_2 representing the number of samples in the kth sample subset. The G statistic is then calculated as the product

$$G = (M)(m) \tag{2-16},$$

where M and m are as defined above. Control limits are set as

$$\begin{cases} UCL = \chi^2_{\left(\frac{p(p+1)}{2}, \frac{\alpha}{2}\right)} \\ LCL = \chi^2_{\left(\frac{p(p+1)}{2}, \frac{1-\alpha}{2}\right)} \end{cases} \tag{2-17},$$

where $\chi^2_{(p,q)}$ is the qth quantile (Levison *et al*., 2002, p. 541).

2.4 OTHER MEASURES OF DISPERSION

There are three other approaches found in the literature that attempt to capture the effect of dispersion on MSPC control charts. These are depicted on Figure 2 as the M (Hayter and Tsui, 1994) chart, the *GS* (Houshmand *et al*., 1997) chart, and the methods of partial and multiple correlation (Golnabi and Houshmand, 1997). While each of these approaches does address the dispersion issue, their directions diverge from that of this research since the objective here is to separate the dispersion from the location to characterize its behavior. Both the M chart and the *GS* chart combine dispersion and location measures into a single, chartable statistic, and are an attempt to address Alt's (1985) recommendation for the development of one control chart for the

simultaneous monitoring of location and dispersion. The *M* chart and *GS* chart are reviewed in Appendix A and Appendix B, respectively.

The similar methods of partial correlation and multiple correlation do directly address changes in the correlation structure itself. However, the process makes iterative comparisons using a method that does not follow the template of the standard Shewhart chart which is the basis for MSPC charting. Thus, the practical difficulties and other implications of charting the generated parameters are not addressed. This algorithmic method is reviewed in Appendix C.

2.5 THE CORRELATION MATRIX AS A MEASURE OF DISPERSION

While three statistics ($|\boldsymbol{S}|$, W_i, and G) have been proposed for tracking changes in the dispersion, there has been no evaluation to compare the relative performances of these statistics. At best, these statistics have been "validated" using data from Jackson (1980) and others. The following section outlines the relationship between $\boldsymbol{R}$ and $\boldsymbol{S}$ as a mathematical justification to comparing the existing statistics ($|\boldsymbol{S}|$, W_i, and G) for monitoring changes in the correlation matrix.

Sample covariance provides a measure of linear association between two variables and is calculated as:

$$s_{ik} = \sum_{j=1}^{n} (X_{ij} - \overline{X}_i)(X_{kj} - \overline{X}_k) \qquad (2\text{-}18),$$

where X_i and X_k are values from two equal-sized (n) sets of variables with means $\overline{X}_i$ and $\overline{X}_k$, respectively, and j is an index. For the population, s_{ik} would be replaced by σ_{ik} and the average

values $\overline{X}_i$ and $\overline{X}_k$ would be replaced by μ_i and μ_k, respectively. The sample correlation coefficient, r_{ik}, is the standardized sample covariance, where the product of the square roots of the sample variances provides the standardization. Thus, the sample correlation coefficient, r_{ik}, can also be viewed as a sample covariance (Johnson & Wichern, p.10).

The sample correlation matrix is comprised of p x p quality characteristics, r_{ik}, each of the form

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} \tag{2-19}$$

where i = 1, 2, . . . , p and k = 1, 2, . . . , p. In this equation, each entry r_{ik} is the Pearson's product moment correlation coefficient for the i^{th} and k^{th} entry in the p x p matrix. If the original values X_{ij} and X_{kj} are replaced by standardized values $\frac{(X_{ij} - \overline{X}_j)}{\sqrt{s_{ii}}}$ and $\frac{(X_{kj} - \overline{X}_k)}{\sqrt{s_{kk}}}$, respectively, then the standardized values are commensurable since both sets are centered at zero and expressed in standard deviation units. The sample covariance of the standardized observations, then, is the sample correlation coefficient, r_{ik} (Johnson & Wichern, p. 10).

Each sample correlation coefficient, r_{ik}, will possess several properties of interest. First, the value of r_{ik} must be between -1 and 1. Note that r_{ik} will equal r_{ki} for all values of i and k, and that any number correlated with itself will have a value of unity. Thus, the diagonals of the matrix $\boldsymbol{R}$ will be equal to one, and the matrix itself will be symmetric. This property means that, unlike the variance-covariance matrix $\boldsymbol{S}$, the correlation matrix $\boldsymbol{R}$ is bounded, allowing a comparison between calculated dispersion statistics.

Second, similar to covariance, the correlation coefficient measures the strength of linear association where r_{ik} = 0 indicates no association, and the sign of r_{ik} indicates direction. For

correlation, $r_{ik} < 0$ indicates a tendency for one value in the pair to be larger than the mean when the other is smaller than its mean and $r_{ik} > 0$ indicates that values in each pair will tend to be either simultaneously large together or small together (Johnson & Wichern, p. 10).

Third, if the values of the i^{th} variable are changed to $y_{ij} = ax_{ij} + b$ and the those of the k^{th} variable are changed to $y_{kj} = cx_{kj} + d$, for $i = 1, 2, \ldots, n$ in both cases, then, provided that a and c have the same sign, there will be no change in r_{ik}.

Fourth, using either n or $n - 1$ (common to avoid bias) as the denominator for s_{ik} results in the same value for r_{ik}.

Despite the desirable qualities of the correlation matrix, $\boldsymbol{R}$, there are some items that remain of concern when using $\boldsymbol{R}$ in place of $\boldsymbol{S}$. Largely these relate to situations where $|\boldsymbol{S}| = 0$ (similarly, then, when $|\boldsymbol{R}| = 0$). When the determinant of a matrix is zero, the matrix is singular and cannot be inverted so that a dispersion statistic cannot be calculated. Note that this does not permit a calculation of the T^2 statistic for the mean vector either. This degenerate case can result from any row of the deviation matrix being expressible as a linear combination of the remaining rows, or in all cases where the number of quality characteristics, p, exceeds the sample size, n. The latter case can be avoided by proper design. In cases where the former situation occurs, Johnson and Wichern suggest:

> that the measurements on some variables be removed from the study as far as the mathematical computations are concerned. The corresponding reduced data matrix will then lead to a covariance matrix of full rank and a nonzero generalized variance. The question of which measurements to remove in degenerate cases is not easy to answer. When there is a choice, one should retain measurements on a (presumed) causal variable instead of those on a secondary characteristic (p. 105).

This was the case with the data analyzed by Holmes and Mergen (1993) using the G chart. The study looked at changes in the distribution of particle size in an industrial process,

where size was categorized as small, medium, or large. The authors state, "The data on three different particle sizes are given . . . only the first two columns are used in the analysis since the total of the percentages is always 100 and the variance-covariance matrix will not invert under these conditions" (p. 622).

Large sample behavior tends to dominate the application of multivariate statistical process control, and the multivariate normal distribution is the distribution usually associated with MSPC processes. Johnson and Wichern note that there are two main reasons for this: (1) for certain natural phenomenon multivariate normal is the population model; and (2) for many statistics, the approximate sampling distribution is multivariate normal (p. 120). Therefore, provided the sample size, n, is large enough, the parent population does not need to be multivariate normal, but must have a mean $\boldsymbol{\mu}$ and finite covariance $\boldsymbol{\Sigma}$ (p. 144). If $\boldsymbol{\Sigma}$ is finite, $\boldsymbol{S}$ and $\boldsymbol{R}$ will be finite as well.

3.0 CORRELATION CONTROL CHART DEVELOPMENT

Since the three statistics, $|\boldsymbol{S}|$, W_i, and G, were developed at different times with different approaches they have not been appropriately compared to one another, nor have they been compared using the in-control *ARL*, the common metric of control charts in which shorter *ARL*s are generally desired (DeVore, 2002). To elaborate, Alt (1985) used a mathematical derivation to develop the W_i statistic and then used a small data set to compare it to the $|\boldsymbol{S}|$ statistic, the latter being a matrix determinant that was assumed to capture dispersion effects (Wilks, 1932). While Alt (1985) had mathematical justification for preferring the W_i statistic to the $|\boldsymbol{S}|$ statistic, closed-form solutions are generally not practical for the comparison of multiple statistics; rather simulations are used, and *ARL*s are determined as the measure of comparison (Montgomery, 1997; Holmes and Mergen, 1998; Levinson *et al.*, 2003). Accordingly, Levinson *et al.* (2003) derived the G statistic and used a simulation producing an *ARL* to show its effectiveness; however, they did not compare the G statistic to the $|\boldsymbol{S}|$ nor the W_i statistic. In the next chapter (Chapter 4), this research uses a terminating sequential simulation to compare the special case of correlation for two of these three statistics, using the in-control *ARL* as the measure of comparison. In this chapter, a mathematical evaluation develops modified versions of the $|\boldsymbol{S}|$ and W_i statistics and also shows that the G statistic cannot be modified for and applied to the correlation case.

3.1 RATIONALE FOR MODIFICATION OF THE COVARIANCE STATISTICS FOR THE APPLICATION TO CORRELATION MATRICES

The three statistics under consideration ($|\boldsymbol{S}|$, W_i, and G) all utilize the generalized sample variance, $|\boldsymbol{S}|$, in some fashion to capture the effects of covariance, a practice that can be traced to Wilks (1932). The concept presented here is that, as a special case of covariance ($\boldsymbol{S}$), correlation ($\boldsymbol{R}$), could be substituted into the equations for each of the statistics and control charts created therefrom. Because of the way the control limits are calculated, which will appear in subsequent sections, if a certain assumption is made these limits do not change when the generalized sample variance of the standardized variables, $|\boldsymbol{R}|$, is substituted for the generalized sample variance of the non-standardized variables, $|\boldsymbol{S}|$, and this is either beneficial or detrimental depending on the statistic considered.

An assumption is required since the generalized sample variance of the standardized variables, $|\boldsymbol{R}|$, only captures a portion of the information contained by the generalized sample variance of the non-standardized variables, $|\boldsymbol{S}|$. In addition to the correlation component, covariance also contains a scale factor component that provides for the standardization and is calculated as a product of the variances, s_{ii}, so that

$$|S| = (s_{11}s_{22}\cdots s_{pp})|R|, \tag{3-1}$$

for a matrix of $i = 1$ to p quality characteristics.

In the context of a control chart, then, the creation of charts based on $|\boldsymbol{R}|$ instead of $|\boldsymbol{S}|$ requires the assumption that the scale factor is the same in the samples as it is in the in-control condition and that out-of-control conditions are generated solely by changes in the correlation matrix, $\boldsymbol{R}$. While this assumption may not always be valid, it is reasonable in MSPC if it is the

correlation component that is of interest and if it is the correlation component that is expected to provide the out-of-control condition. That is, the scale factor component is not being ignored; rather the concentration is on the correlation component's contribution to covariance. Such matrices occur in a number of processes ranging from chemical fractionating to anthropometric data to financial data and meet the requirements of an Exchange Structure which will be discussed in Chapter 5. Furthermore, the assumption of scale factor stability can be tested once an out-of-control condition is detected by the correlation control chart by also viewing the control chart for the covariance that utilizes $|\boldsymbol{S}|$, since $|\boldsymbol{S}|$ does not capture only the effects of correlation (Johnson & Wichern, 103). This approach is analogous to performing a linear regression and then examining the residuals to verify normality because that assumption is sometimes inconvenient to assess *a priori*. When the generalized sample variance is decomposed as shown in equation 3-1, the effects of the scale factor and the correlation are separated, as will be shown geometrically later, so that the correlation control charts become companions to the covariance control charts.

Because different correlation structures are not detected by the generalized sample variance of the non-standardized samples, and different correlation structures are detected by the generalized sample variance of the standardized samples, this suggests the following in regard to these control charts:

1. If the correlation chart is in-control and the covariance chart is in-control, then the process is in control.
2. If the correlation chart is out-of-control and the covariance chart is in-control, then the process *may be* out-of-control due to correlation.

3. If the correlation chart is in-control and the covariance chart is out-of-control, then the process *may be* out-of-control due to the scale factor.
4. If the correlation chart is out-of-control and the covariance chart is out-of-control, then the process is out-of-control due to both the correlation and the scale factor.

Thus, even though the correlation control chart could be seen mathematically as a special case of the covariance chart, the covariance chart would in practical applications be considered the companion chart to the correlation chart even if the primary objective is to detect changes in the correlation matrices generated from the process data.

3.2 RELATION OF $|R|$ TO $|S|$

Since the three covariance statistics being considered use $|\boldsymbol{S}|$ as part of their calculations, it is instructive to see the relationship between $|\boldsymbol{S}|$ and $|\boldsymbol{R}|$. Although equation 3-1 explains this relation mathematically, geometry of the bivariate or trivariate cases provides a visual explanation. It is extendable by induction for $p > 3$, where p represents the number of quality characteristics since, even if the p vectors are n-dimensional, p vectors can span no more than a p-dimensional space. If $p = 2$, a plane is defined. If $p = 3$, a volume is defined. Similarly, when extended for $p > 3$ the spaces become hypervolumes (specifically parallelotopes) that are difficult to visualize but nonetheless valid (Anderson, 1984).

Johnson & Wichern (1988), and similarly Anderson (1984), develop the notion of a deviation vector in (p =) 3-space that is useful in illustrating the relation of $|\boldsymbol{S}|$ to $|\boldsymbol{R}|$.

Define a vector

$$y_i'\left(\frac{1}{\sqrt{n}}A\right)\left(\frac{1}{\sqrt{n}}A\right)=\frac{x_{iA}+x_{iB}+\cdots+x_{in}}{n}A=\bar{x}_i A \tag{3-2}$$

where $\bar{x}_i$ is the sample mean and n is the dimension of an n x 1 vector $A'=[1,1,\ldots,1]$, so that $\frac{1}{n}y_i'A$ corresponds to the multiple of A that gives the projection of y_i onto the line determined by A. For each y_i, the geometric decomposition would be as shown in Figure 3.

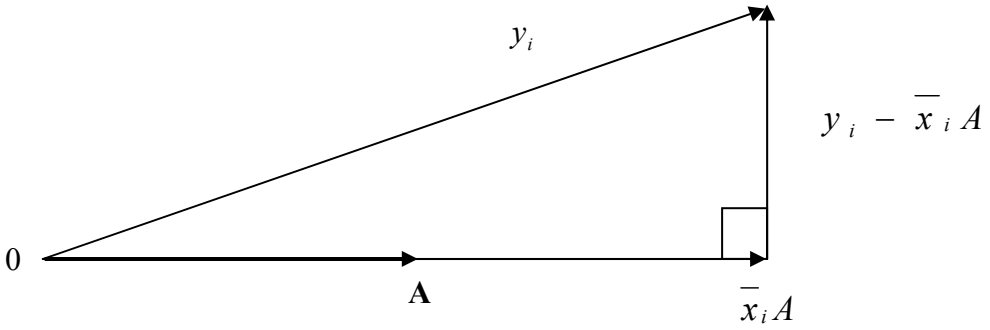

Figure 3 Decomposition of deviation vector

The deviations of the measurements on the i^{th} variable from their respective sample means, then, are given by the difference vector e_i

$$e_i=y_i-\bar{x}_i A=\begin{bmatrix}x_{iA}-\bar{x}_i\\ x_{iB}-\bar{x}_i\\ \vdots\\ x_{in}-\bar{x}_i\end{bmatrix}, \tag{3-3}$$

where the other variables are as defined previously.

The vector A has been defined purposefully to lie anywhere in n-space, as the axes of the deviation vectors, e_i, can now be translated to the origin without affecting their lengths nor their orientations. A typical view showing three deviation vectors is shown in Figure 4.

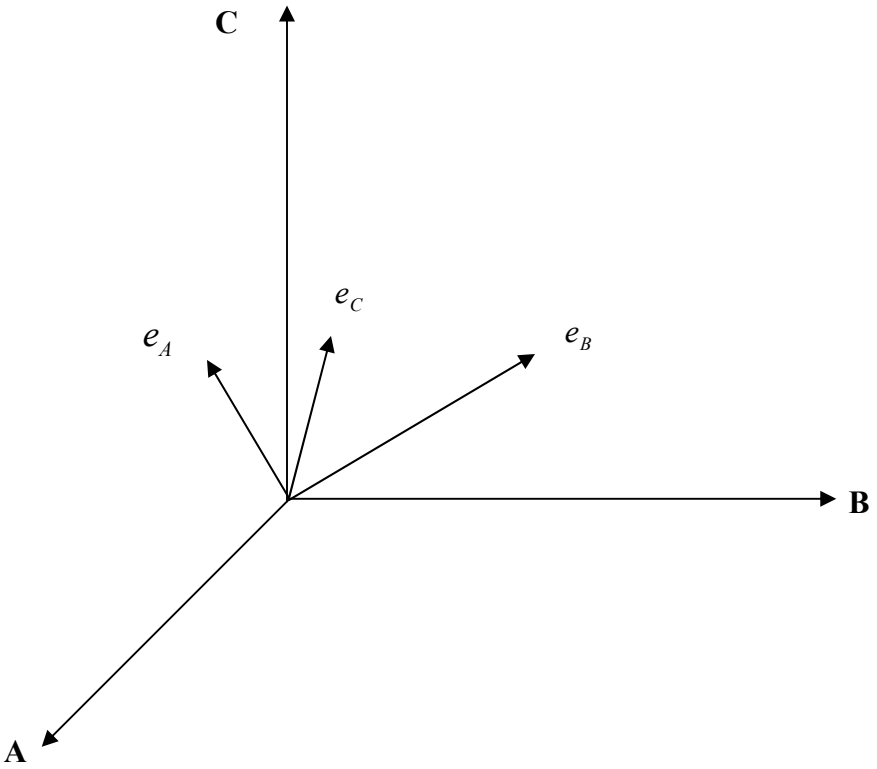

Figure 4 Deviation vectors on A-B-C axes

The squared lengths of the deviation vectors, which are proportional to the variance of the measurements on the i^{th} variable, are equivalent to the sum of the squared deviations, meaning that the lengths are proportional to the standard deviations. The cosine of the angle between pairs of deviation vectors is the correlation coefficient given by the equation

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos(\theta_{ik}) \qquad (3\text{-}4)$$

where r_{ik} denotes the correlation coefficient, s_{ik} is a variance, and the square roots of s_{ii} and s_{kk} are standard deviations. Figure 5 shows the geometric interpretation, in (p = 2) dimensions,

between any two deviation vectors $e_1 = y_1 - \bar{x}_1 A$ and $e_2 = y_2 - \bar{x}_2 A$. Since the height of the trapezoidal region formed by the deviation vectors is $L_{e_1} \sin(\theta)$, the area bounded by the vectors and their projections is $L_{e_1} L_{e_2} \sqrt{1 - \cos^2(\theta)}$. Johnson & Wichern show that, by using substitutions and extending to a multidimensional case by induction, that the volume generated in *n* space for *p*

deviation vectors is given by

$$|\boldsymbol{S}| = (n-1)^{-p} (\text{volume})^2. \tag{3-5}$$

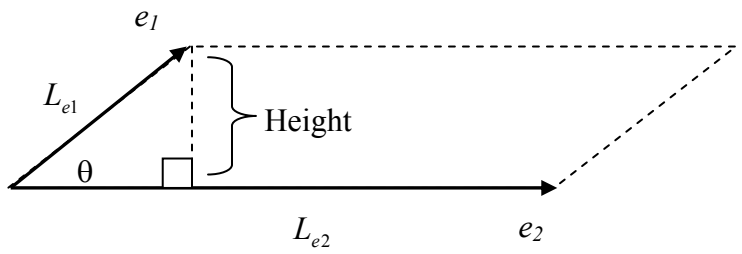

Figure 5 Trapezoidal region formed by deviation vectors

Looking at this in three dimensions, shown in Figure 6, it is evident that two components affect the volume—the length of the sides, and the angles between edges.

The correlation corresponds to the angle between deviation vectors and the lengths of the edges correspond to the scale factor (usually expressed in standard deviation units). Thus, in the covariance case, either the scale factor or the correlation may affect the volume and, therefore, the sample generalized variance, $|\boldsymbol{S}|$.

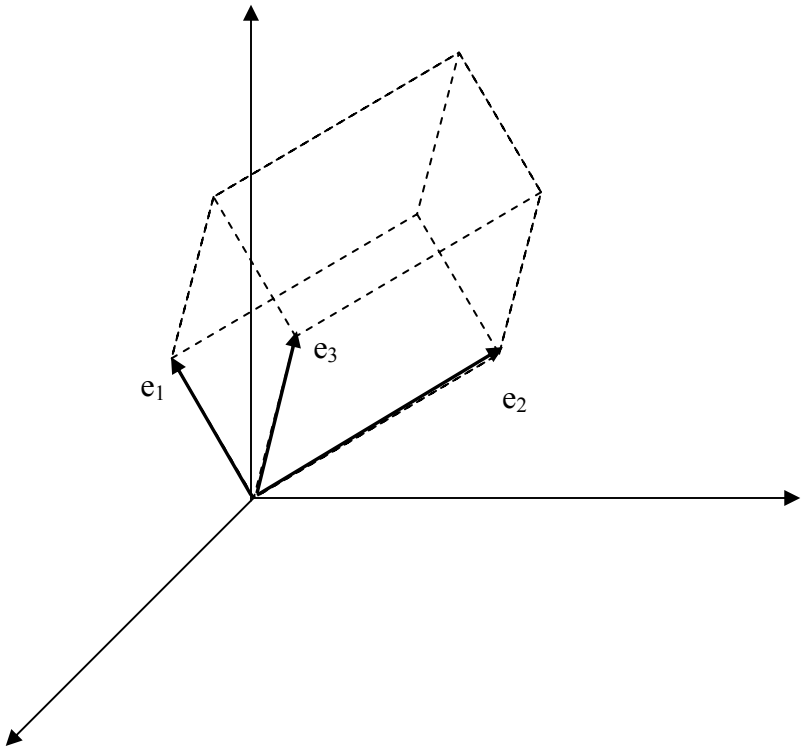

Figure 6 Parallelotope in Three Dimensions

When the covariance matrix is standardized, so that the standardized covariance, or correlation matrix, $\boldsymbol{R}$, is obtained, all of the vector lengths are equivalent and only the angle between vectors effects the volume. The geometric representation is the same as that of Figure 6, with all the e_i being of equal lengths. Assume, then, that the generalized sample variance of the standardized sample is used instead of the non-standardized version in calculation of the three covariance statistics ($|\boldsymbol{S}|$, W_i, and G), and that these statistics behave similarly as hypothesized. Then replacing $|\boldsymbol{S}|$ with $|\boldsymbol{R}|$ in the covariance statistics would lead to control charts that detect when correlation has gone out-of-control assuming the scale factor has remained constant between the in-control condition and subsequent samples. As stated, this assumption of

constant scale factor should be investigated by comparing the correlation control chart with the covariance control chart when an out-of-control condition for correlation is detected.

3.3 COVARIANCE STATISTICS APPLIED TO CORRELATION[1]

Since correlation is a special case of covariance, the statistics used for MSPC monitoring of covariance should apply to the monitoring of correlation. To investigate the plausibility of correlation as a separate component, it is necessary that the scale factor does not change between the in-control condition and the samples—an assumption that must be checked by also using the covariance control chart once the correlation control chart detects an out-of-control condition. In doing so, the control limits for the correlation case are generally commensurate with those for the original covariance statistics and this is shown in subsequent sections. As Table 1 indicates, there is a slight modification of the control limits when considering $|\boldsymbol{R}|$ instead of $|\boldsymbol{S}|$, but no modification when considering W_R instead of W_i. These limits will be derived in the following sections. Issues with the G statistic will be addressed in Section 3.3.3.

[1] Thanks to Dr. Murat C. Testik, Ph.D. for reviewing the derivation of control limits and other calculations presented in Chapter 3. Dr. Testik received his doctorate from Arizona State University where he studied MSPC with Dr. Douglas Montgomery.

Table 1 The Statistics and Their Control Limits

Covariance Statistic	Correlation Statistic	Control Limits
$\|\boldsymbol{S}\|$	Not applicable	$\left[\|\Sigma\|\frac{\left(\chi^2_{2n-4,1-(\alpha/2)}\right)^2}{[4(n-1)^2]}, \|\Sigma\|\frac{\left(\chi^2_{2n-4,\alpha/2}\right)^2}{[4(n-1)^2]}\right]$
Not applicable	$\|\boldsymbol{R}\|$	$\left[\|\rho\|\frac{\left(\chi^2_{2n-4,1-(\alpha/2)}\right)^2}{[4(n-1)^2]}, \|\rho\|\frac{\left(\chi^2_{2n-4,\alpha/2}\right)^2}{[4(n-1)^2]}\right]$
W_i	W_R	$[0, \chi^2_{(p(p+1)/2,\,\alpha}]$

3.3.1 Control Limits for $|\boldsymbol{R}|$

When $p = 2$, $|S|$ is distributed exactly as chi-squared. However, for p > 2, Alt (1973) suggests using Anderson's approximation as

$$P\left[\left|\Sigma_0\right|\left(1-z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right)\le|S|\le\left|\Sigma_0\right|\left(1+z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right)\right]=1-\alpha \tag{3-5}$$

where Σ_0 is the in-control covariance matrix, S is the sample covariance matrix, p denotes the number of quality characteristics, n is the sample size, and $z_{\alpha/2}$ is from the standard normal distribution. Then the control limits for the $|S|$ chart are

$$UCL=\left|\Sigma_0\right|\left(1+z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right) \tag{3-6}$$

and

$$LCL = \left|\Sigma_0\right|\left(1 - z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right) \tag{3-7}$$

where the variables are as noted above. Since $\Sigma_0 = V^{1/2}\rho V^{1/2}$, where V represents the diagonal matrix of variances so that $V^{½}$ is the diagonal matrix of standard deviations and ρ represents the correlation matrix for the in-control condition, substituting into the first equation gives

$$P\left[\left|V^{1/2}\rho V^{1/2}\right|\left(1 - z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right) \le \left|\hat{V}^{1/2} R \hat{V}^{1/2}\right| \le \left|V^{1/2}\rho V^{1/2}\right|\left(1 + z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right)\right] = 1-\alpha \tag{3-8}$$

where R is the sample correlation matrix. Now, if A and B are k x k square matrices—which ***S*** and ***R*** have to be—and it has been shown that

$$\left|AB\right| = \left|A\right|\left|B\right| \tag{3-9}$$

the equation now becomes

$$P\left[\left|V^{1/2}\right|\left|\rho\right|\left|V^{1/2}\right|\left(1 - z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right) \le \left|\hat{V}^{1/2}\right|\left|R\right|\left|\hat{V}^{1/2}\right| \le \left|V^{1/2}\right|\left|\rho\right|\left|V^{1/2}\right|\left(1 + z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right)\right] = 1-\alpha \tag{3-10}$$

Since the determinants are scalars and it is assumed that, under the in-control conditions, the standard deviation matrices for the sample and population are equivalent, and sample sizes are sufficiently large, dividing by the common factor leaves

$$P\left[\left|\rho\right|\left(1-z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right)\le\left|R\right|\le\left|\rho\right|\left(1+z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right)\right]=1-\alpha \tag{3-11}$$

meaning that the control limits for the $|R|$ chart would be

$$UCL=\left|\rho\right|\left(1+z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right) \tag{3-12}$$

and

$$LCL=\left|\rho\right|\left(1-z_{\alpha/2}\sqrt{\frac{2p}{n-1}}\right), \tag{3-13}$$

so that a control chart could be constructed for correlation based on the $|\boldsymbol{S}|$ covariance control chart.

However, a significant issue with the limits shown in Equations 3-12 and 3-13 is very evident. The determinant of a correlation matrix is bounded as [0,1] so that the upper and lower limits of Equations 3-12 and 3-13, respectively, will rarely be exceeded, especially when the magnitude of the correlation exceeds 0.4. This result is an indicator that the performance of the $|\boldsymbol{R}|$ statistic is questionable (or that correlation, thus measured, cannot be viewed as a major contributor to the covariance).

3.3.2 Control Limits for W_R

Use of the W_i statistic (Alt, 1973) is equivalent to making repeated hypothesis tests of the form

$$H_0 : \Sigma = \Sigma_0 \text{ versus } H_1 : \Sigma \neq \Sigma_0$$

where Σ and Σ_0 are the population covariance and matrix of in-control covariance values, respectively.

The statistic is a ratio of the maximum likelihood estimators for the multivariate normal distribution (μ, Σ) in the following sets:

$$\Omega = \{(\mu, \Sigma) : -\infty < \mu < \infty, \Sigma \_ is \_ positive \_ definite\}$$

$$\omega = \{(\mu, \Sigma) : -\infty < \mu < \infty, \Sigma = \Sigma_0\}.$$

This implies that the mean is unspecified and the covariance matrix is specified, to be used in the likelihood ratio statistic to derive the control statistic. Note that this is important since the statistic now becomes invariant to changes in the mean vector.

It has been shown (Anderson, 1984) that the maximum likelihood estimators for the multivariate normal case are:

$$\hat{\mu}_{\Omega} = \bar{x} \qquad \hat{\Sigma}_{\Omega} = \frac{1}{n} A \tag{3-14a, b}$$

$$\hat{\mu}_{\omega} = \bar{x} \qquad \hat{\Sigma}_{\omega} = \Sigma_0 \tag{3-15a, b}$$

where $\bar{x}$ is the vector mean, n is the sample size, and $A = (n-1)S$ so that the likelihood functions based on the multivariate normal distribution are

$$L(\hat{\Omega}) = \left(\frac{2\pi}{n}\right)^{-pn/2} |A|^{-n/2} e^{-pn/2}, \tag{3-16}$$

and

$$L(\omega) = (2\pi)^{-pn/2} |\Sigma_0|^{-n/2} e^{-\frac{1}{2} tr(\Sigma_0^{-1} A)}, \tag{3-17}$$

(where *tr* is the trace operator) giving a maximum likelihood estimate

$$\Lambda(x) = \frac{1}{n} e^{pn/2} \left|\Sigma_0^{-1} A\right|^{n/2} e^{-\frac{1}{2} tr(\Sigma_0^{-1} A)}. \tag{3-18}$$

Note here that the maximum likelihood of the covariance matrix is not an unbiased estimator. Taking the log likelihood, the statistic becomes

$$W_i = -2\ln(\Lambda) = -pn + pn\ln n - n\ln\left(\frac{|A|}{|\Sigma_0|}\right) + tr(\Sigma_0^{-1} A) \tag{3-19}$$

which is rejected any time its value exceeds $\chi^2\left(\frac{p(p+1)}{2}, \alpha\right)$, an expected result for the log likelihood. This is the asymptotic general result, since the specification of Σ_0 requires $\frac{p(p+1)}{2}$ independent elements to be specified.

To approach this statistic from the standpoint of correlation in which the goal is repeated tests of the hypotheses

$$H_0 : \rho = \rho_0 \text{ versus } H_1 : \rho \neq \rho_0,$$

one could calculate the maximum likelihood ratio for a new Ω and ω defined in terms of the correlation. However, since the multivariate normal distribution requires introduction of the variance for full characterization, the result will be a test of the covariance.

If, instead, we look at the third and fourth terms of the W_i statistic, we see that they include the ratio $\frac{|A|}{|\Sigma_0|}$, which is equivalent to $\frac{|(n-1)S|}{|\Sigma_0|}$, or $(n-1)^p \frac{|S|}{|\Sigma_0|}$, where S is replacing Σ in equation 3-19, and depicts a scalar $(n\text{-}1)^p$ multiplied by the ratio of the determinants of the sample covariance matrix to the in-control covariance matrix. To convert this ratio for the in-control correlation matrix to the ratio of the determinants of the sample correlation matrix, $\frac{|R|}{|\rho_0|}$, requires making the assumption that the scale factors for both cases are equivalent, an assumption that has been noted must be checked once an out-of-control condition is indicated. The statistic, by substitution, then becomes

$$W_R = -2\ln(\Lambda) = -pn + pn\ln n - n\ln\left((n-1)^p \frac{|R|}{|\rho_0|}\right) + (n-1)tr(\rho_0^{-1}R), \qquad (3\text{-}20)$$

so that a control chart can be constructed for correlation, W_R, based on the W_i covariance control chart.

In addition to being intuitive, there is a justification for this substitution since the maximum likelihood of the correlation occurs at the same point as the maximum likelihood of the covariance.