Georgia Journal of Science

Volume 67 *No. 2 Scholarly Contributions from the Membership and Others*

Article 1

2009

Principle Components for Diagnosing Dispersion in Multivariate Statistical Process Control

Terrance E. Murphy
terrance.murphy@yale.edu

Mary McShane-Vaughn

Kwok Leung-Tsui

PRINCIPAL COMPONENTS FOR DIAGNOSING DISPERSION IN MULTIVARIATE STATISTICAL PROCESS CONTROL

Terrence E. Murphy, PhD[1]
Mary McShane-Vaughn, PhD[2]
Kwok Leung-Tsui, PhD[1]

[1]Georgia Institute of Technology, Atlanta, GA
[2]Southern Polytechnic State University, Marietta, GA

Corresponding Author
Terrence E. Murphy
6 Hunting Ridge
Hamden, CT
Email: terrence.murphy@yale.edu
telephone: 203-737-2295
fax: 203-785-4823

ABSTRACT

We provide an easily implemented procedure to help data analysts systematically diagnose which quality characteristics may be driving the dispersion of a multivariate process out of control. Multivariate statistical process control (MSPC) commonly uses Hotelling's T^2 statistic to indicate when a multivariate observation goes out-of-control. Several techniques currently exist that accurately diagnose which specific variables are driving the T^2 statistic out-of-control. For subgroups of independently and identically distributed multivariate normal observations, we advocate decomposing the overall T^2 into independent T^2 statistics for separate monitoring of location and dispersion. We propose a procedure based on principal components to diagnose the specific variables responsible for driving subgroup dispersion out-of-control. The procedure is demonstrated on a publicly available data-set.

Keywords: multivariate, statistical process control, principal components, diagnosis of dispersion

INTRODUCTION

It is increasingly clear that new methods of diagnosing the dispersion of multivariate processes are needed. (1) The purpose of this article is to present a principal component based procedure for diagnosing which specific variable(s) in multivariate statistical process control (MSPC) are driving the process dispersion out of control. Well regarded sources in the literature that summarize the state of MSPC (2, 3, 4, 5) indicate that one of the test statistics most commonly used to monitor a multivariate process is Hotelling's T^2 statistic. Although principal components have been used in the past to diagnose which variables are driving the T^2 statistic out of control (2, 3, 4, 5) , the efficacy of principal component based (PC-based) diagnosis has been contingent upon how well these components approximate a physically interpretable latent factor. Kourti and MacGregor (6) show that under multivariate normality the normalized scores of the principal components can accurately diagnose the causal variables regardless of their physical meaning. In contrast, Mason and Young (5) diagnose the responsible variables by an orthogonal decomposition of the T^2 statistic not based on principal components.

Data analysts are increasingly monitoring multivariate processes that are subject to shifts in scale as well as location. When monitoring rational subgroups of multivariate observations, the techniques of Mason and Young (5) and Kourti and MacGregor (6) have limited diagnostic capability because they can't discern between these two different kinds of shifts, i.e. scale and location. Mason and Young (5) state that their decomposition procedure does not discretely identify the specific variables responsible for shifting location versus those responsible for driving dispersion out-of-control. Kourti and MacGregor (6) did not demonstrate how their procedure might be applied to diagnosis of multivariate dispersion.

Our contribution is to show how the procedure of Kourti and MacGregor (6) can be extended to systematically identify variables driving the T^2 statistic for subgroup dispersion out-of-control. We demonstrate the diagnostic potential of our procedure on a data-set from Fuchs and Kenett (4) containing subgroups of data whose significant T^2 values are directly attributable to out-of-control subgroup dispersion. To our knowledge there is no such procedure based on a simply implemented decomposition of Hotelling's T^2 statistic. We hope this procedure will be useful to data analysts who monitor multivariate processes by providing a systematic method of investigating which variables may be driving process variance out of control.

The paper is organized as follows. In Materials and Methods we show how the PC-based procedure of Kourti and MacGregor (6) can be directly applied to diagnose shifts in subgroup location and extended to diagnose shifts in subgroup dispersion. In Results we demonstrate the diagnostic procedure for subgroup dispersion on data from Fuchs and Kenett (4). In Discussion we present concluding remarks.

MATERIALS AND METHODS

Diagnosis of Shifts in Hotelling's T^2

In the practice of statistical process control, it is preferable to collect rational subgroups of multivariate observations whenever possible. Subgroups yield more reliable process information than individual vectors as well as an estimate of the correlation structure within the subgroup. Research including that of Hawkins (7), Kourti and MacGregor (6) and Mason and Young (5) can be used to diagnose the variables which drive the subgroup's overall T^2 out of control. However these do not typically differentiate between the variables shifting location versus those driving higher dispersion. In order to distinguish which variables cause which type of shift, we first decompose this statistic into two independent parts representing subgroup location and dispersion.

Decomposition of Hotelling's T^2 into Location and Dispersion

Consider a rational subgroup k consisting of the n individual multivariate data-points $\mathbf{Y}_{ki}$ where $i = 1 \ldots n$ and each $\mathbf{Y}_{ki}$ is of dimension r. The overall subgroup T^2 for subgroup k, i.e., T^2_{0k}, is the sum of the T^2_{ki} of the $i = 1 \ldots n$ individual multivariate data-points that make up subgroup k. T^2_{0k} is also known as the Lawley-Hotelling trace statistic (Lawley (8) and Hotelling (9)) and asymptotically has a $\chi^2_{(rn)}$ distribution (Jackson (10)).

Jackson (11) discussed the following decomposition of Hotelling's T^2_{0k}:

$$T^2_{0k} = T^2_{Mk} + T^2_{Dk}\,, \qquad \textit{(equation 1)}$$

where T^2_{Mk} can be used to test whether the sample mean of this subgroup has shifted away from the estimated population mean and is defined as:

$$T^2_{Mk} = n(\bar{\mathbf{Y}}_k - \hat{\boldsymbol{\mu}})^T \hat{\Sigma}^{-1} (\bar{\mathbf{Y}}_k - \hat{\boldsymbol{\mu}})\,, \qquad \textit{(equation 2)}$$

where $\bar{\mathbf{Y}}_k$ is the vector whose components are the subgroup averages of the n multivariate observations in subgroup k. The distribution of T^2_{Mk} is asymptotically $\chi^2_{(r)}$, the same as that of an individual observation, i.e. T^2_{ki} (Jackson (11)).

In contrast T^2_{Dk} can be used to test whether the variance within this subgroup is significantly greater than historical variance and is defined as:

$$T^2_{Dk} = \sum_{i=1}^{n} (\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)^T \hat{\Sigma}^{-1} (\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k), \qquad \textit{(equation 3)}$$

where $\mathbf{Y}_{ki}$ the are the $i = 1 \ldots n$ multivariate observations making up subgroup **k**. The distribution of T^2_{Dk} is asymptotically $\chi^2_{r(n-1)}$ where r is the number of quality characteristics in each multivariate observation and n the number of observations in the subgroup (Jackson (11)).

Whereas T^2_{0k} contains all the information in the subgroup, Jackson (12) argues that it is of little diagnostic value since when significant, the analyst must immediately ascertain whether it is T^2_{Mk}, the multivariate analog of the $\bar{x}$ chart, or T^2_{Dk}, the multivariate analog of the r chart, that is responsible.

Applying the Diagnostic Procedure of Kourti and MacGregor

In this section we show how the PC-based procedure of Kourti and MacGregor (6) diagnoses the specific quality characteristics shifting subgroup location and extend it to do the same for increased dispersion. As in their procedure, we assume that the multivariate observations are independently and identically distributed as multivariate normal. Since the PC's are derived from decomposition of the estimated covariance matrix $\hat{\Sigma}$, we first re-express $\hat{\Sigma}^{-1}$ in terms of the eigenvectors and eigenvalues of $\hat{\Sigma}$.

Since $\hat{\Sigma}$ is positive definite, it can be expressed as $\hat{\Sigma} = \mathbf{u}\Lambda\mathbf{u}^T$, where $\mathbf{u}$ is an r by r orthogonal matrix whose columns are the eigenvectors of $\hat{\Sigma}$ and Λ is an r by r diagonal matrix containing the pertinent eigenvalues. The inverse of $\hat{\Sigma}$ can be expressed in similar manner as

$$\hat{\Sigma}^{-1} = \mathbf{u}\Lambda^{-1}\mathbf{u}^T = [\mathbf{u}\Lambda^{-1/2}]\,[\mathbf{u}\Lambda^{-1/2}]^T \qquad \textit{(equation 4)}$$

The covariance matrix of subgroup mean $\bar{\mathbf{Y}}_k$ is $\frac{\hat{\Sigma}}{n}$. Its inverse can be expressed as

$$\left(\frac{\hat{\Sigma}}{n}\right)^{-1} = \mathbf{v}\Lambda_{\bar{\mathbf{Y}}}^{-1}\mathbf{v}^T = \left[\mathbf{v}\Lambda_{\bar{\mathbf{Y}}}^{-\frac{1}{2}}\right]\left[\mathbf{v}\Lambda_{\bar{\mathbf{Y}}}^{-\frac{1}{2}}\right]^T, \qquad \textit{(equation 5)}$$

where the columns of $\mathbf{v}$ are the eigenvectors of $\frac{\hat{\Sigma}}{n}$ and the diagonal elements of $\Lambda_{\bar{\mathbf{Y}}}$ are the corresponding eigenvalues.

Diagnosing Shifts in Subgroup Location

Mason et al (3) state that due to its use of principal components, the approach of Kourti and MacGregor (6) is particularly useful for large and ill-conditioned data sets. Although their procedure was demonstrated for a single observation, it is easily shown that the procedure works equally well to diagnose which specific variables are driving a subgroup average out of control. Plugging (*equation 5*) into the decomposition of subgroup T^2_{Mk} in *equation 2* yields

$$\begin{aligned} T^2_{Mk} &= \left[\mathbf{v}^T\Lambda_{\bar{\mathbf{Y}}}^{-\frac{1}{2}}\left(\bar{\mathbf{Y}}_k - \hat{\boldsymbol{\mu}}\right)\right]^T \left[\mathbf{v}^T\Lambda_{\bar{\mathbf{Y}}}^{-\frac{1}{2}}\left(\bar{\mathbf{Y}}_k - \hat{\boldsymbol{\mu}}\right)\right] \\ &= \sum\nolimits_{p=1}^{r}\left[\frac{\mathbf{v}_p{}^T\left(\bar{\mathbf{Y}}_k - \hat{\boldsymbol{\mu}}\right)}{\pi_p^{1/2}}\right]^2 \end{aligned} \qquad \textit{(equation 6)}$$

which we recognize as the sum of squares of normalized scores based on the vector of differences between the subgroup mean vector $\bar{\mathbf{Y}}_k$ and the estimated population mean vector $\hat{\boldsymbol{\mu}}$, where the individual columns of $\mathbf{v}$, i.e. the eigenvectors $\mathbf{v}_p$, and eigenvalues $\pi_p^{1/2}$ for $p = 1 \ldots r$, are derived from the covariance matrix of $\bar{\mathbf{Y}}_k$, i.e $\frac{\hat{\Sigma}}{n}$.

For a specific subgroup k, we define the p^{th} normalized score of location ($NSL_{k,p}$) as

$$NSL_{k,p} = \left[\frac{\mathbf{v}_p{}^T\left(\bar{\mathbf{Y}}_k - \hat{\boldsymbol{\mu}}\right)}{\pi_p^{1/2}}\right] \qquad \textit{(equation 7)}$$

where the indices k and p respectively index the specific subgroup and score being considered. The $NSL_{k,p}$ are approximately standard normal so any value greater than or equal to three is statistically different from its population mean at an $\alpha = 0.0027$. A simple bar-chart of the $NSL_{k,p}$ quickly identifies which are out of control. Variable contributions to the $NSL_{k,p}$ (vc$NSL_{k,p,j}$), are defined as follows,

$$\text{vc}NSL_{k,p,j} = \frac{v_{p,j}\left(\bar{Y}_{k,j} - \hat{\mu}_j\right)}{\pi^{1/2}}, \qquad (\textit{equation 8})$$

where $v_{p,j}$ is the element of eigenvector p corresponding to the individual variable j and $\bar{Y}_{k,j}$ is the average of the values of the individual variable j from the $i = 1 \dots n$ observations in subgroup k. We can now write out T^2_{Mk} as

$$T^2_{Mk} = \sum_{p=1}^{r} \left[\frac{\mathbf{v}_p^T\left(\bar{\mathbf{Y}}_k - \hat{\boldsymbol{\mu}}\right)}{\pi_p^{1/2}}\right]^2 \qquad (\textit{equation 9})$$

$$= \sum_{p=1}^{r} \left[\sum_{j=1}^{r} \left\{\frac{v_{p,j}\left(\bar{Y}_{k,j} - \hat{\mu}_j\right)}{\pi_p^{1/2}}\right\}\right]^2$$

which shows its relation between the individual variable contributions and the squared $NSL_{k,p}$. For $n > 1$, the procedure of Kourti and MacGregor (6) diagnoses the variables responsible for driving subgroup mean out-of-control by following these steps:

1. A subgroup with an out-of-control value of T^2_{Mk} is detected at some level of α.
2. Plot a bar-chart of the $NSL_{k,p}$.
3. For each significant $NSL_{k,p}$, plot a bar-chart of the (vc$NSL_{k,p,j}$).
4. Investigate those variables making a large contribution of the same sign as the statistically significant $NSL_{k,p}$.

Extending Kourti and MacGregor to Diagnose Shifts in Subgroup Dispersion

Plugging (*equation 4*) into the decomposition of subgroup T^2_{Dk} in (*equation 3*) yields

$$T^2_{Dk} = \sum_{i=1}^{n} \left(\left[\mathbf{u}_p^T \Lambda^{-1/2}(\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)\right] \left[\mathbf{u}_p^T \Lambda^{-1/2}(\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)\right]\right) \qquad (\textit{equation 10})$$

$$= \sum_{i=1}^{n} \sum_{p=1}^{r} \left[\frac{\mathbf{u}_p^T(\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)}{\lambda_p^{1/2}}\right]^2$$

which we recognize as the sum of squares of normalized scores calculated from the vector of differences between each individual observation $\mathbf{Y}_{ki}$ and the subgroup mean vector $\bar{\mathbf{Y}}_k$. We define scores of dispersion ($ScD_{k,p,i}$) as

$$ScD_{k,p,i} = \mathbf{u}_p^T(\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k) \qquad (\textit{equation 11})$$

where k denotes the subgroup, p the principal component, and i the individual observation within the subgroup. Within any subgroup the average of $\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k$ is a vector of zeroes, so we assume that the $ScD_{k,p,i}$ are approximately normal as follows,

$ScD_{k,p,i} \sim N(0,\lambda_p)$.

Because the $ScD_{k,p,i}$ are normalized by dividing by the square root of the associated eigenvalue, we define normalized scores of dispersion ($NScD_{k,p,i}$) as follows,

$$NScD_{k,p,i} = \left[\frac{\mathbf{u}_p{}^T(\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)}{\lambda_p^{1/2}}\right], \qquad (equation\ 12)$$

which are approximately standard normal, *i.e.*,

$NScD_{k,p,i} .\sim N(0,1)$

We can now re-write (*equation 10*) as the sum of the squares of the $NScD_{k,p,i}$ from the n observations in subgroup k as follows,

$$T^2{}_{Dk} = \sum_{i=1}^{n}\sum_{p=1}^{r}\left[\frac{\mathbf{u}_p{}^T(\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)}{\lambda_p^{1/2}}\right]^2 \qquad (equation\ 13)$$

$$\sum_{i=1}^{n}\sum_{p=1}^{r}\left[NScD_{k,p,i}\right]^2$$

$$\sum_{p=1}^{r}\sum_{i=1}^{n}\left[NScD_{k,p,i}\right]^2$$

Based on the approximate standard normality of the $NScD_{k,p,i}$, we can show that within subgroup k, the sum of the squares of the n observations of $NScD_{k,p,i}$ are approximately distributed as $\chi^2_{(n-1)}$, i.e.,

$$\sum_{i=1}^{n}\left[NScD_{k,p,i}\right]^2 \sim \chi^2{}_{(n-1)} \,. \qquad (equation\ 14)$$

An outline of the proof of (*equation 14*) is provided elsewhere (13). Within that proof it is shown that the sum of squared $NScD_{k,p,i}$ is a constant multiple of the sample variance of the $NScD_{k,p}$.

This means that simply plotting the sums of squared $NScD_{k,p,i}$ reveals which of the $NScD_{k,p}$ have large sample variance. We proceed to decompose those $NScD_{k,p}$ with large sample variance into contributions from individual variables. For each observation $\mathbf{Y}_{ki}$ in subgroup k, individual variable contributions to the normalized scores of subgroup dispersion (vc$NScD_{k,p,i,j}$) are defined as follows:

$$\text{vc}NScD_{k,p,i} = \frac{u_{p,j}\left(Y_{k,i,j} - \bar{Y}_{k,j}\right)}{\lambda_p^{1/2}}, \qquad (equation\ 15)$$

where the subscript k indicates the subgroup, p the specific score, i the specific observation and j the individual variable. $Y_{k,i,j}$ is the value of variable j in observation $\mathbf{Y}_{ki}$ and $\bar{Y}_{k,j}$ the average of the values of variable j within subgroup k. This definition of vc$NScD_{k,p,i,j}$ enables the following expression of (*equation 10*):

$$T^2_{Dk} = \sum_{p=1}^{r}\sum_{i=1}^{n}\left[NScD_{k,p,i}\right]^2$$

$$=\sum_{p=1}^{r}\sum_{i=1}^{n}\left[\sum_{j=1}^{r}\frac{u_{p,j}\left(Y_{k,i,j}-\bar{Y}_{k,j}\right)}{\lambda_p^{1/2}}\right]^2 \qquad (equation\ 16)$$

which shows the relation between individual variable contributions and the values of the $NScD_{k,p,i}$. In our procedure, for each specific value of p where the sum of squared $NScD_{k,p,i}$ is statistically significant, we will plot the standard deviations of the variable contributions to the $NScD_{k,p,i}$. For subgroups of $n > 1$, we propose the following extension of Kourti and MacGregor (6) to diagnose the variables responsible for driving subgroup dispersion out-of-control:

1. A subgroup with an out-of-control value of T^2_{Dk} is detected at some level of α.
2. Plot a bar-chart of all the $\sum_{i=1}^{n}\left[NScD_{k,p,i}\right]^2$.
3. Look for values of $\sum_{i=1}^{n}\left[NScD_{k,p,i}\right]^2$ exceeding the critical value of $\chi^2_{(n-1)}$ at α.
4. For each $NScD_{k,p}$ whose value of $\sum_{i=1}^{n}\left[NScD_{k,p,i}\right]^2$ is statistically significant, calculate the standard deviations of the $vcNScD_{k,p,i,j}$ from the n observations of $NScD_{k,p,i}$ in subgroup k.
5. Plot a bar chart of the standard deviations computed in the previous step.
6. Investigate those variables with the largest standard deviations of $vcNScD_{k,p,i,j}$.

RESULTS

Demonstration and Comparison of the Proposed Procedure

To demonstrate and contrast techniques in MSPC, Fuchs and Kenett (4) employ a data-set they call Case 1 consisting of seventy observations of six dimensions each. This data-set is divided into thirty five subgroups of two apiece where the first fifteen subgroups are used to generate historical (Phase I) estimates of the mean vector and the covariance matrix. The last twenty are used as the Phase II data-set for comparison against the historical Phase I values. Figures 1, 2 and 3 respectively plot T^2_{Ok}, T^2_{Mk}, and T^2_{Dk} values for the first seventeen of the twenty distinct subgroups in Phase II, where critical values at $\alpha = 0.01$ for the three statistics are indicated by the horizontal lines of asterisks. The upper control limits (UCL) are calculated as follows (14, 15):

UCL for T^2_{Ok} : $\chi^2_{(30,\ \alpha=0.01)} = 50.89$ based on 30 datapoints from Phase I

UCL for T^2_{Mk} (14): $\frac{6(15+1)(2-1)}{15(2-1)-6+1}F^{\alpha=0.01}_{6,10} = (9.6)*(5.39) = 51.74$

UCL for T^2_{Dk} (15) : (2-1) $\chi^2_{(6,\ \alpha=0.01)}$ = 16.81 for subgroups of n = 2

Figures 1, 2, and 3 emphasize the point made by Jackson (12) that a significant value of T^2_{0k} typically requires further inquiry as to whether location, dispersion or both are driving the subgroup out-of-control.

Figure 1: To^2 for Phase II Subgroups of Fuchs-Kenett Case 1

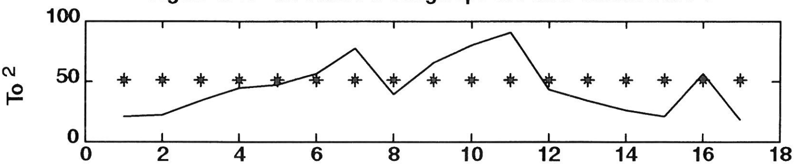


Figure 2: Tm^2 for Phase II Subgroups of Fuchs-Kenett Case 1

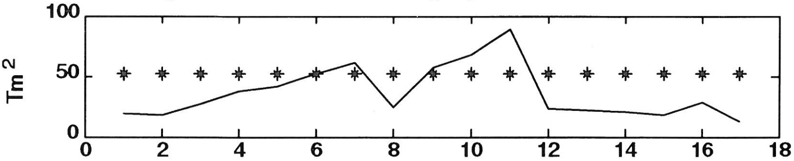
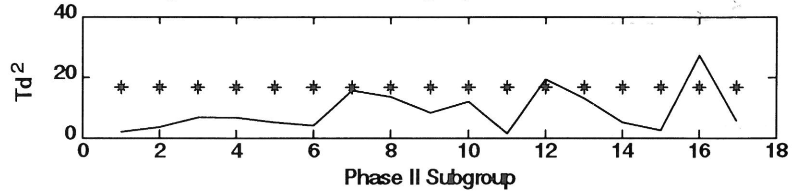


Figure 3: Td^2 for Phase II Subgroups of Fuchs-Kenett Case 1



We will focus on subgroups 11 and 16 whose T^2_{0k} are significant at α = 0.01. Subgroup 11 has a significant T^2_{Mk} and non-significant T^2_{Dk} while the converse is true for subgroup 16.

Demonstration of the Diagnostic Procedure of Kourti and MacGregor

Figures 4 and 5 are the results of applying the procedure of Kourti and MacGregor (6) for diagnosing which variables are causing a shift in location of subgroup 11, where the critical value at α = 0.0027 is 3.00.

Figure 4: Normalized Scores of Location (NSL) from Subgroup 11

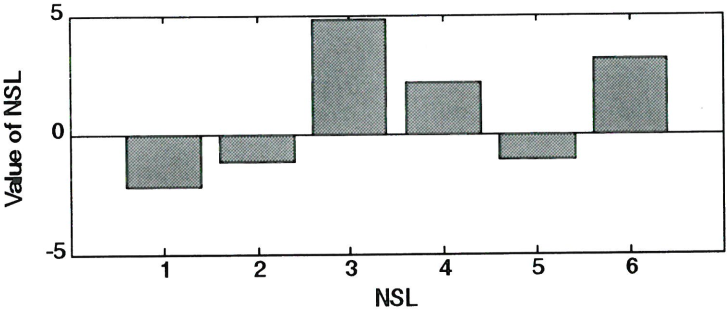
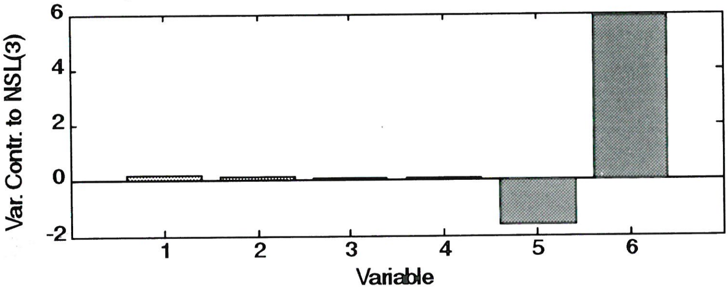


Figure 5: Variable Contributions to NSL(3) from Subgroup 11



In Figure 4 the third score from subgroup 11, i.e. $NSL_{11,3}$, is most significant followed by $NSL_{11,6}$. Because the $NSL_{k,p}$ are approximately standard normal, any value greater than or equal to 3 is highly significant. Figure 5 shows that variable 6 makes the largest same sign contribution to $NSL_{11,3}$. By adjusting the value of variable 6 in one of the data points from subgroup 11 closer to its Phase I mean, the value of T^2_{M11} decreased from 89.09 to 40.87, well below the critical value of 51.74.

Figures 6 and 7 are the results of applying our procedure for diagnosing dispersion to subgroup 16. In Figure 6 the critical value at $\alpha = 0.01$ is 6.63. Note that because in this example, where $n = 2$, the sum of squared $NScD_{16,p,i}$ is simply the sample variance of the $NScD_{16,p,i}$.

Figure 6: Sample Variances of Normalized Scores of Dispersion (NScD) from Subgroup 16

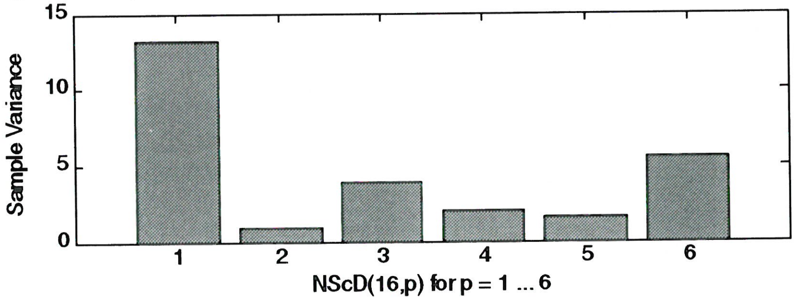


Figure 7: Std Dev (SD) of Variable Contr. to NScD(1) from Subgroup 16

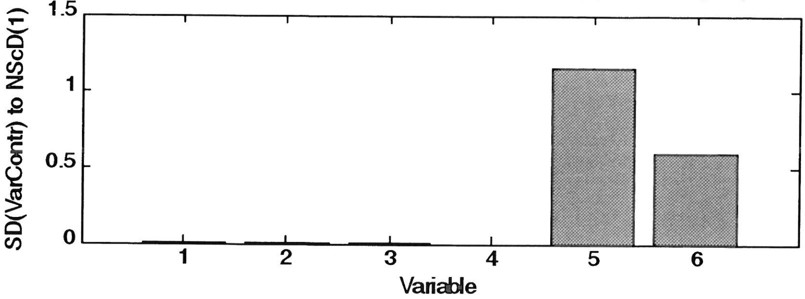


Figure 6 shows that for $p = 1 \dots 6$, i.e. the six different $NScD_{16,p}$, the sum of squared values corresponding to $NScD_{16,1}$ is of highest significance. Figure 7 shows that the contributions of variables 5 and 6 have the largest variance in decreasing order. By adjusting the values of variables 5 and 6 closer to their subgroup averages in one of the observations making up subgroup 16, the value of T^2_{D16} decreased from 27.19 to 11.92, well below the critical value of 16.81.

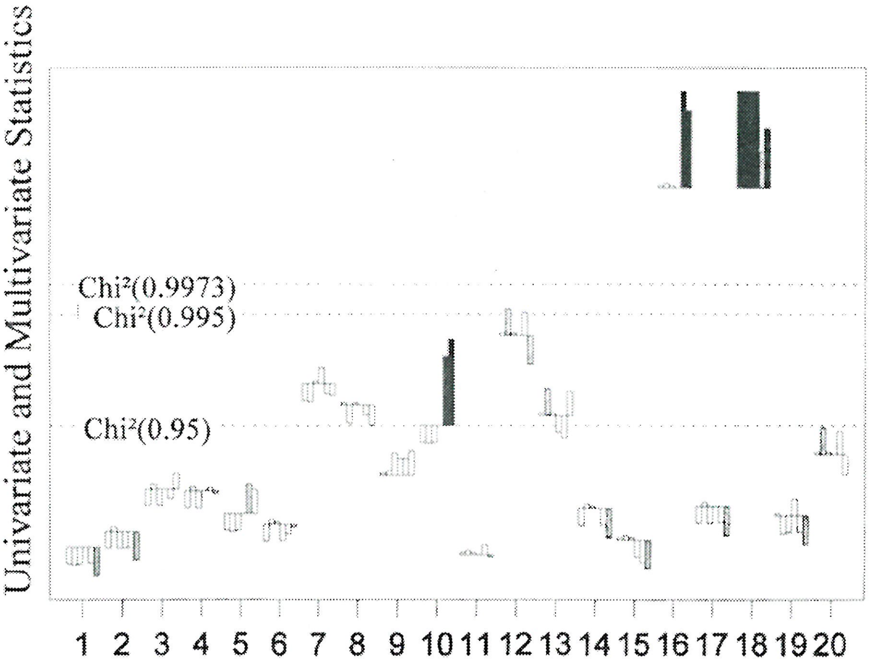
Comparing Techniques for Diagnosis of Subgroup Dispersion

Of the works referenced in this paper, only Fuchs and Kenett (4) demonstrate a technique for identifying the individual variables that drive a particular subgroup's variance out of control. They employ a graphical diagnostic tool called the multivariate profile (MP) chart which can be constructed for separate diagnosis of shifts in location and variation. MP charts were introduced by Fuchs and Benyamini (16) and position the baseline of a miniature bar-chart of scaled deviations at the vertical magnitude of the respective T^2_{Mk} or T^2_{Dk} statistic. The miniature bar-chart for location, which is vertically positioned at the magnitude of the subgroup's T^2_{Mk} value, displays the scaled deviation of each individual variable's subgroup mean with respect to standard values of location. The miniature bar-chart for dispersion, which is vertically positioned at the magnitude of the subgroup's T^2_{Dk} value, displays the scaled deviation of each individual variable's subgroup dispersion with respect to standard values of dispersion.

The miniature bar-charts are essentially individual variable charts of measures of location and dispersion which do not consider correlation and are not directly related to the decomposition of the relevant T^2 statistic. The MP chart's simultaneous presentation of the multivariate T^2 statistic and bar-charts of univariate scaled deviation provide an informative snapshot of subgroup behavior which often facilitates quick identification of the suspect variables.

Figure 8 is a reproduction of page 130 of Fuchs and Kenett (4) which plots the MP charts for dispersion for the last twenty subgroups of the Case 1 data-set. Notice that a curve formed by connecting the baselines of each subgroup's MP dispersion chart mimics the shape of the T^2_{Dk} chart in Figure 3.

Figure 8: MP Dispersion Charts for Last 20 Subgroups of Case 1 from Fuchs-Kenett



In Figure 8 subgroups 16 and 18 are out of control at an α of 0.0027 and subgroups 7, 8, 10, 12 and 13 are out of control at an α of 0.05. The MP dispersion chart suggests that individual variables with scaled deviation of greater magnitude be investigated first. The degree of darkness of the individual bars denotes increasing levels of significance. Notice that the miniature bar-charts of subgroups 10, 16 and 18 each have individual variables whose deviations are very dark and tower above the others. This means that those individual variables, namely variables 5 and 6 in subgroups 10 and 16 and variables 1 through 4 in subgroup 18, are clearly indicated as responsible for driving the subgroup dispersion out of control. For these three subgroups, where the subgroup dispersion of individual variables is much higher than standard values, our procedure diagnoses the same variables as the MP chart for dispersion. We verified that for subgroups 10, 16, and 18, the same causal variables identified by both procedures were responsible by adjusting one or more of their values closer to the subgroup mean before re-calculating T^2_{Dk}. This resulted in T^2_{Dk} values well below the critical values for all reasonable α levels.

For subgroups 7, 8, 12 and 13 there are no individual variables in the MP chart for dispersion whose scaled deviations differ extremely from standard values. Nonetheless the miniature bar-charts suggest first investigating those variables with scaled deviation of greater magnitude. To contrast our

PC-based diagnostic procedure for subgroup dispersion with the MP dispersion charts, we restrict our discussion to subgroup 8. The value of subgroup 8 is significant at $\alpha = 0.05$ with a critical value of 12.59. The MP dispersion chart of subgroup 8 draws attention to variables 2 and 6 due to their larger magnitudes of scaled deviation from standard values of dispersion. In Figures 9 and 10 we present the diagnosis from our procedure.

Figure 9: Sample Variances of Norm. Scores of Dispersion from Subgroup 8

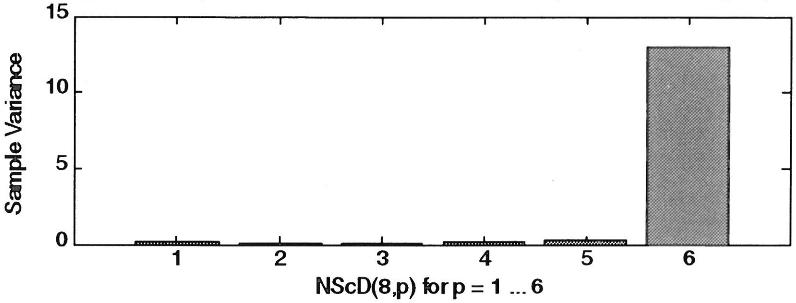
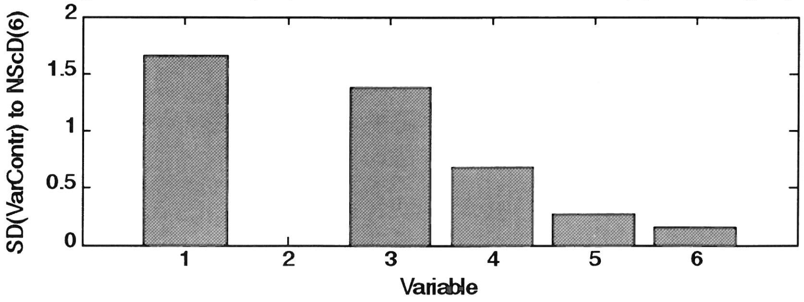


Figure 10: Std Dev (SD) of Variable Contributions to NScD(6) from Subgroup 8



In Figure 9 the critical value at $\alpha = 0.01$ is 6.63. Notice that of the six different $NScD_{8,p}$ for $p = 1 \ldots 6$ within subgroup 8, only the sum of squared values of $NScD_{8,6,i}$ is significant at $\alpha = 0.01$. Figure 10 shows that the contributions of variables 1, 3 and 4 to $NScD_{8,6}$ have the largest variation in decreasing order. To evaluate the diagnostic accuracy of the proposed PC-based procedure, we modify the original subgroup 8 data to reduce the differences between the identified variables and their subgroup means. Since variables 1 and 3 are diagnosed by our procedure, we halve the distance between them and their subgroup means in one of subgroup 8's observations. This reduces the T^2_{D8} value from 13.67 to 4.69, well below the critical value of 12.59.

For this same subgroup the MP chart for dispersion points us toward variables 2 and 6. From Figure 10 we see that the standard deviation of variable 2 in subgroup 8 is zero, which explains why in the MP chart for dispersion the scaled deviation for this variable is a large negative value. The large negative value means that in this subgroup the second variable's deviation is much lower than its standard value. Although this makes variable 2 a non-factor in driving T^2_{D8} out of control, it may indicate another problem since a sample variance of zero is suspicious. Halving the distance between the values of variable 6 and their subgroup mean reduces the T^2_{D8} value from 13.67 to 13.25, still above the critical value of 12.59. The comparison of diagnoses between the principal components based procedure and the MP chart indicates that for this specific subgroup, the PC-based procedure for diagnosing subgroup dispersion is more informative.

DISCUSSION

There are several reliable techniques for identifying the individual variables responsible for driving the T^2 value of a multivariate observation out-of-control. Less research has been published regarding the diagnosis of subgroups of multivariate observations, which are prone to shifts of scale as well as location. The MP charts of Fuchs and Benyamini (16) are a helpful graphical instrument for investigating the potentially causal variables of rational subgroups whose location and or dispersion have shifted away from standard values. In their discussion, Fuchs and Kenett (4) indicate that the scaled deviations of the MP dispersion chart must be interpreted in concert with the correlation structure of the variables for accurate diagnosis. In contrast, the PC-based procedure introduced here is directly related to the decomposition of the portion of Hotelling's T^2 corresponding to dispersion, thereby explicitly integrating the correlation structure of the individual variables.

We have extended the PC-based technique described in Kourti and MacGregor (6) to diagnose the causal variables of a particular subgroup's shift in dispersion. This technique is simply implemented and, owing to its incorporation of the correlation structure of the quality characteristics being monitored, is more informative than the corresponding MP chart in some cases.

Because the procedure is examining a large number of relationships, there is an obvious concern for multiple comparisons. This procedure does not easily adjust for multiple comparisons. Kourti and MacGregor (6) address this by stating that once the T^2 statistic has gone out of control, a deviation at the chosen level of significance has already been detected. They posit that because the univariate charts of the normalized scores serve primarily as guides for pinpointing the errant variables, a precise adjustment of p-values for multiple comparisons is not warranted. We support this reasoning and assert that if T^2_M or T^2_D go out of statistical control, the data analyst is most concerned with identifying the probable cause as soon as possible. Even in the case of many observed factors, the procedure described here provides data analysts with a directed approach to help them more quickly identify variables potentially linked to shifts in scale.

The need for new methods of diagnosing the dispersion of multivariate observations has recently been noted in the quality engineering literature (1). Given the dearth of user friendly techniques available for diagnosing statistical processes whose dispersion has shifted out of control, we feel the utility of this approach more than offsets its difficulty in rigorously accounting for multiple comparisons.

ACKNOWLEDGMENTS

T.E. Murphy and K.-L. Tsui gratefully acknowledge support from the National Science Foundation, grant DMI-0100123. T.E. Murphy's work was also supported by OMED Educational Services at Georgia Institute of Technology. K.-L. Tsui's work was also supported by The Logistics Institute - Asia Pacific in Singapore.

REFERENCES

1. Yeh AB, Lin DKJ, McGrath RN: Multivariate control charts for monitoring covariance matrix: a review. Journal of Quality Technology and Quantitative Management 3: 415-436, 2006.
2. Lowry CA, Montgomery DC: A review of multivariate control charts. IIE Transactions 27: 800-810, 1995.
3. Mason RL, Champ CW, Tracy ND, Wierda SJ, Young JC: Assessment of multivariate process control techniques. Journal of Quality Technology 29(2): 140-162, 1997.
4. Fuchs C, Kenett RS: "Multivariate Quality Control Theory." NY, Marcel Dekker, p. 130, 1998.
5. Mason RL, Young JC: "Multivariate Statistical Process Control with Industrial Applications." Alexandria VA and Philadelphia PA, ASA-SIAM, p. 1-30, 2002.
6. Kourti T, MacGregor JF: Multivariate SPC methods for process and product monitoring. Journal of Quality Technology 28(4): 409-428, 1996.
7. Hawkins DM: Regression adjustment for variables in multivariate quality control. Journal of Quality Technology 25(3): 170-182, 1993.
8. Lawley D: A generalization of Fisher's z-test. Biometrika 30: 180-187, 1938.
9. Hotelling H: A generalized T test and measure of multivariate dispersion. Proceedings of the 2nd Berkeley Symp. Of Math. Stat. and Prob 1: 23-41, 1951.
10. Jackson JE: "A User's Guide to Principal Components." NY, Wiley, p. 125, 1991.
11. Jackson JE: Principal components and factor analysis: part II – additional topics related to principal components. Journal of Quality Technology 13(1), 46-58, 1981.
12. Jackson JE: Multivariate quality control. Commun. Stat. Meth. 14(11): 2657-2688, 1985.

13. Murphy TE: Multivariate quality control using loss-scaled principal components. PhD at Georgia Institute of Technology, p. 235, 2004. Web Document: http://etd.gatech.edu/theses/available/etd-11222004-122326/unrestricted/murphy_terrence_e_200412_phd.pdf
14. Fuchs C, Benyamini Y: Multivariate profile charts for statistical process control. Technometrics 36: 71, 1994.
15. Fuchs C, Benyamini Y: Multivariate profile charts for statistical process control. Technometrics 36: 40, 1994.
16. Fuchs C, Benyamini Y: Multivariate profile charts for statistical process control. Technometrics 36: 182-195, 1994.