Appl. Math. and Comp. Sci., 1998, Vol.8, No.4, 775–792

CONTRIBUTION PLOTS: A MISSING LINK IN MULTIVARIATE QUALITY CONTROL

PAIGE MILLER*, RONALD E. SWANSON**
CHARLES E. HECKLER***

Most multivariate quality control techniques involve plotting and analysis of a set of surrogate variables, such as T^2, scores and residuals. The number of these surrogate variables can be considerably smaller than the number of original variables. However, it is often difficult to determine the source of a problem when the process is identified as being out of control by one of the surrogate variables. We have called this difficulty the "missing link" in multivariate quality control. In this article, contributions and contribution plots are introduced as a simple method to correct this problem and enhance the interpretation of the multivariate results, exploration of data and identification of special causes.

1. Introduction

One aspect of quality control requires the detection of special causes, or out-of-control situations, including outliers, level shifts, trends or patterns. Consider a process that is measured by many variables, sometimes highly correlated with each other. Multivariate quality control techniques use the entire data set, taking into account the correlation between the process variables, to detect special causes in the process. Once a special cause is found, the engineers and operators need to understand the data in such a way that the actual problem can be diagnosed. If the actual problem can then be eliminated from the process, improvement of the quality of the process will occur. This paper presents a new tool for helping diagnose a problem found by using multivariate statistical methods. Jackson (1991, p. 21) identified four goals of multivariate quality control:

1. A single answer should be available to the question: "Is the process in control?"
2. An overall Type I error should be specified.

* Eastman Kodak Company, 7th Floor/Building 6, Kodak Park, Rochester, NY 14652–4608, e-mail: paigem@kodak.com.
** Eastman Kodak Company, 6th Floor/Building 6, Kodak Park, Rochester, NY 14652–4612, e-mail: rswanson@kodak.com.
*** Eastman Kodak Company, 9th Floor/Building 83, Kodak Research Labs, Rochester, NY 14650–2208, e-mail: checkler@kodak.com.

3. The procedure should take into account the relationship among the variables.

4. Procedures should be available to answer the question: "If the process *is* out of control, what is the problem?"

Jackson goes on to say: "Condition 4 is much more difficult than the other three, particularly as the number of variables increases. There usually is no easy answer to this, although the use of PCA (Principal Components Analysis) may help. The other three conditions are much more straightforward." Wierda (1994) echoes the same sentiment, after reviewing numerous multivariate SPC procedures: "The most important open question from a practical point of view is how to detect the variables that caused the out-of-control signal."

We call Jackson's fourth goal the *missing link*. Miller and Swanson (1993) introduced a simple solution called *contributions*, and a simple plot, called a *contribution plot* to address this issue. The contribution plot provides the missing link that lets us interpret multivariate statistical information in terms of Jackson's fourth goal. They are exploratory in nature and help us interpret the special causes in our data by providing a clear link back to the original variables which might have caused the out-of-control signal. This paper expands upon the original idea. Since the introduction of contribution plots, other authors such as MacGregor and Kourti (1995), Hopkins *et al.* (1995) and Kourti and MacGregor (1996) have successfully applied contribution plots. Another method of addressing the issue of identifying the variables that are out of control was suggested by Hayter and Tsui (1994) — contribution plots will be compared to the Hayter and Tsui approach later. Also, Fuchs and Benjamini (1994) introduced a new type of graphical chart to detect the variables that are causing an observation to be out of control. In addition to specialized software needed to create the graphics in Fuchs and Benjamini (1994), Kourti and MacGregor (1996) describe drawbacks to the Fuchs and Benjamini approach.

2. Description of the Application

We shall introduce the idea of contribution plots via an example. Our application involves the monitoring of a photographic emulsion manufacturing process. In this process, salt, silver and other chemicals are added to a kettle at the appropriate times and then are mixed at a certain speed and temperature. Data are regularly collected on a variety of process variables, including flows, pressures, temperatures, pH, mixer speeds, etc. The data are available for analysis at the completion of a batch, and thus some form of statistical process control (SPC) is appropriate to determine if the process is in-control or out-of-control. If the process is out-of-control, the engineer needs to know what parts of the process actually had problems, thus motivating our development of contribution plots. From this information, the search for the actual physical cause begins, and then the cause can be eliminated from future batches.

In this example, we shall use a history of 230 batches, all made on the same piece of equipment under the same operating conditions. We have 27 different process measurements available.

3. The T^2 Control Chart

The first investigative tool is the T^2 control chart (see e.g. Jackson, 1991). We note that there are other multivariate charting schemes available, in place of the T^2 control chart, such as multivariate CUSUM charts (Crosier, 1988; Woodall and Ncube, 1985), or multivariate EWMA charts (Lowry *et al.*, 1992). These other methods also suffer from the same drawback as T^2, namely that it is not easy to detect which process variable(s) is the cause of the out-of-control signal. Also of interest here is a multivariate charting approach based upon regression adjustment of variables (Hawkins, 1993).

Assuming multivariate normality of the data, the T^2 chart satisfies Jackson's first three criteria. The formula for T^2 for observation i is:

$$T_i^2 = \mathbf{x}_i \mathbf{S}^{-1} \mathbf{x}_i' \tag{1}$$

where $\mathbf{x}_i$ is a (row) vector representing the process variable measurements for batch i, and $\mathbf{S}$ is the variance-covariance matrix of the data. (Usually, T^2 calculations are done on the centered data, and also the data can be optionally scaled so that each variable has a variance of 1. Without loss of generality, references in this article to data $\mathbf{x}_i$ in both formulas and text will refer to the centered and optionally scaled data.) In practice, $\mathbf{S}$ is based on a set of data where the process was in control and $\mathbf{x}_i$ will be some future observation being evaluated. For simplicity, both $\mathbf{S}$ and $\mathbf{x}_i$ come from the same data.

It is not clear how best to determine which subset of the data represent this "in-control" operation of the process when 27 variables are involved. One idea is to do PCA on the entire data, find the outliers, and then use the remaining data for a multivariate statistical process control scheme. The data shown in this example is the initial PCA on all batches; this analysis found so many interesing and unexpected results that it was judged to be very valuable by itself, and so issues concerning selection of the proper data set or "robustness" of PCA to outliers were deferred until a later point in time. Devlin *et al.* (1981) discuss methods of robust PCA estimation.

It is well known that T^2 can be computed from the PCA scores (see e.g. Jackson, 1991). We choose to reduce the number of dimensions going into the T^2 calculation by selecting only the first a PCA dimensions. Thus, T^2 for observation i is

$$T_i^2 = \sum_{d=1}^{a} t_{id}^2 / \lambda_d \tag{2}$$

where the PCA scores t in dimension d have variance λ_d, which is the d-th largest eigenvalue of $\mathbf{S}$. (See (Jackson, 1991) for details on the computation of the scores t_{id}.) There are many methods of choosing the value of a, the number of dimensions, e.g. see (Jackson, 1991, Section 2.8; Wold, 1978). This choice of a is critical because it lets us monitor our complete process with fewer variables while being less sensitive to random (sensor) noise.

PCA has the property that the first dimension explains the most variance of any linear combination of the variables; the second dimension explains the most variance

of any linear combination of the variables that is perpendicular to the first dimension; and so on. Given limited resources, selecting the first a dimensions corresponds to monitoring the phenomena in our process causing the most variability, while ignoring some of the smaller sources of variability; this is a natural choice for engineers trying to improve process performance.

In this example, dimension reduction is plausible. We believe that although we collect 27 different process measurements, there are not 27 independent phenomena going on. The PCA dimensions are a representation of the process variation in a smaller dimensional space, taking into account the correlation between the process variables.

It should be noted that Wise *et al.* (1990) have shown that manufacturing process changes that result in shifts in the dynamics that are driving the process can be detected via multivariate quality control. Wise *et al.* (1990) have specifically shown that an arbitrary dynamic linear time invariant state-space model can always be transformed so that the states are directly related to the PCA scores. In addition, they emphasized that multivariate quality control is most effective when the process has significantly more measurements than states (a situation that occurs in our application).

The T^2 control chart is shown in Fig. 1, based upon 10 PCA dimensions, along with the 95% control limit. Although we suspect that our data is not multivariate normally distributed, these limits are displayed anyway for a "rough" guideline. From this chart, we can see numerous batches which were out-of-control, along with stretches where the process was in-control (e.g. approximately batches 110–140). The engineer, upon seeing a particular batch is out-of-control, needs to know what process variables caused T^2 to indicate this out-of-control situation.

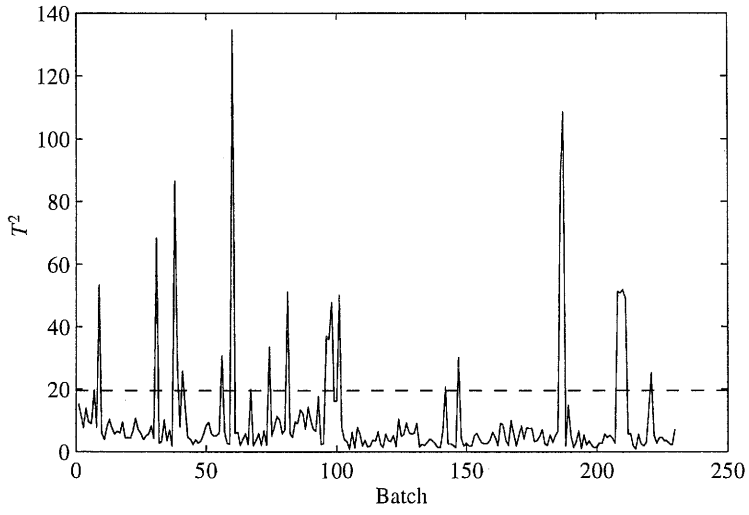

Fig. 1. T^2 control chart based upon 10 dimensions, with 95% control limit.

Let us examine further batch 208. We can see clearly that the T^2 value is above the 95% control limit. Examination of the scores t_{id} would indicate that dimensions 1 and 2 are where the problem might be found. For engineers, the notion that score 1 might be the cause is not actionable: knowing that score 1 is out-of-control is not sufficient information to begin an investigation of the possible physical cause of the problem, unless the dimension has a substantive interpretation (a condition that cannot be guaranteed). More information is needed, relating the scores to the original variables that contribute to the calculation of the scores. Thus we introduce the concept of contribution to scores, which forms the (previously missing) link between an out-of-control signal on a multivariate chart and the original variables that caused the out-of-control signal.

We can write the scores as the weighted sum of the data. The loadings for each dimension are the weights. Thus

$$t_{id} = \sum_{j=1}^{k} x_{ij} p_{jd} \tag{3}$$

where p_{jd} is the loading for variable j in dimension d, and there are k process variables used in the calculations. Thus we can decompose t_{id} into k terms $x_{ij}p_{jd}$ for $j = 1, \ldots, k$. These k terms are the contributions to the score t_{id}. The contribution plot is a bar chart of these k contributions, scaled as indicated below. The contributions for this batch in dimension 1 and the loadings in dimension 1 are shown in Fig. 2.

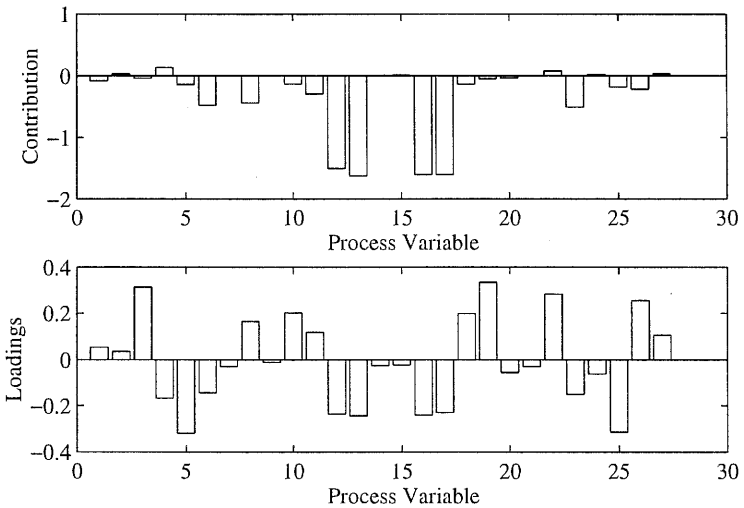

Fig. 2. Contributions to batch 208 and loadings in dimension 1.

Conceptually, contributions are different than loadings. Loadings represent variability across the *entire* data set. Contributions represent the particular process

variables that were unusual *for a given batch*. Our goal is not to interpret the loadings; instead, we use contributions to help us interpret events that are identified as special causes by querying the underlying data.

A practical difference between contributions and loadings occurs when some of the process variables have a value close to zero, even though those same variables may have large loadings. This is illustrated in Fig. 2. The loadings suggest that variables 3, 5, 19 and 25 are important—these variables related to silver concentration difference from setpoint during the start of the batch (3), difference in silver concentration from the start to the end of phase 1 of the batch (5), silver concentration standard deviation during phase 2 of the batch (19) and total amount of salt delivered (25). The contributions tell a different story, with variables 12, 13, 16 and 17 having the biggest contributions. All four of these variables refer to problems with silver or salt pressures (12 and 13 relating to silver pressure and 16 and 17 relating to salt pressure). Thus interpreting the loadings would potentially detect a different process problem for this batch than actually occurred, which was a pressure problem in both the salt and silver delivery systems. Investigation found this problem to be caused by the installation of a different type of valve.

We point out that control charts confirm that variables 12, 13, 16 and 17 are indeed out-of-control in this batch, while control charts for the variables with the four biggest loadings do not indicate an out-of-control condition. We recommend the use of univariate control charts as a confirmatory practice following a diagnosis by use of contribution charts. Although this is not necessary (and sometimes ineffective), it seems to be an important psychological reinforcement to engineers and operators who might be uncomfortable with multivariate statistical calculations.

We note that there is no definition of what constitutes a big contribution. This has been left up to the judgment of the observer, just as in past efforts to interpret the loadings, the selection of large loadings was often subjective. Research needs to be done to quantify which contributions are significantly different from zero. Since we have already identified the batch as being out-of-control, the purpose of the contribution plots is to suggest to the engineer where to begin the investigation. In the vast majority of the cases, information found via contribution plots has turned out to be practically significant, in spite of the absence of formal statistical tests.

4. Residual Contributions

Recall that our definition of T^2 does not use all possible dimensions. Thus we have some loss of information for each batch, which can be quantified into a residual measure Q_i for batch i, defined as

$$Q_i = (\mathbf{x}_i - \hat{\mathbf{x}}_i)(\mathbf{x}_i - \hat{\mathbf{x}}_i)' \tag{4}$$

where $\hat{\mathbf{x}}_i$ is the (row) vector of predicted values of the (centered and scaled) data for batch i based upon a Principal Component dimensions.

A large Q value can be interpreted that the observation does not follow the covariance structure estimated by $\mathbf{S}$. This can happen when either a sensor fails, or the process shifts resulting in a new covariance structure for the process.

Sensor problems can be serious enough to warrant immediate attention if the number reported by the sensor is used for some type of feedback or feedforward control. If the sensor problems occur or the process shifts, Q will increase to reflect this problem. When this is the case, the engineer should attempt to eliminate the new process phenomena and return the process to its operating state observed in the training set. If this is not possible or desirable, and the new process phenomena will continue, then a new $\mathbf{S}$ matrix for computation of T^2 is needed.

But again the question arises—just like it did for T^2—if Q is out-of-control, which process variables caused it to be out? Again, we can compute contribution plots for Q. There are k elements in $\mathbf{x}_i - \hat{\mathbf{x}}_i$, and the squares of these k values are plotted as bars in a contribution plot for Q. Examples of a Q control chart and contributions for an out-of-control batch are shown in Figs. 3 and 4, respectively. For batch 101, which has the largest Q value, the process variables which are unusual and which contribute the most to the large Q are variables 22 (standard deviation of salt flow during phase 2) and 25 (total amount of salt delivered), and there may be some secondary contribution from variable 26 (final silver concentration average). Note that here there are no indications that anything was amiss during phase 1 of the batch and that the biggest problems came from the improper behavior of the salt delivery system. The engineering investigation discovered a sensor problem (typical for large Q values) affecting the feedback control.

5. Comparisons with Other Common Approaches

Before continuing with the example, we discuss some of the potential advantages of using contribution charts, compared to other methods. Specifically, why do we not simply perform the investigation using 27 control charts of the process variables? What is gained by the multivariate approach? We present several advantages of using multivariate statistics with contribution plots:

1. Multivariate statistics can detect multivariate outliers, while control charts cannot. Specifically, T^2 and/or Q can flag points that would not be detected by 3 s limits on control charts. In this data there were two such batches, although in any given data set multivariate outliers may be more or less frequent.
2. Univariate control charts of many variables give misleading information (i.e. false signals), unless the limits for each are appropriately widened. This is impractical when there are a great many variables. Multivariate methods avoid this difficulty by controlling the type I error at acceptable levels in a straightforward manner. For example, suppose a process had ten variables that were approximately mutually independent (possibly included in a larger set of correlated variables) that were all in control with 2 s control limits. The probability that any of these 10 variables is within its ± 2 s limits is approximately 0.95. The probability that all are within the limits is $.95^{10} \approx 0.60$ (assuming independence of all 10 variables). Thus there is a 40% chance that at least one variable will exceed its ± 2 s limits, when, in truth, all variables are in control.

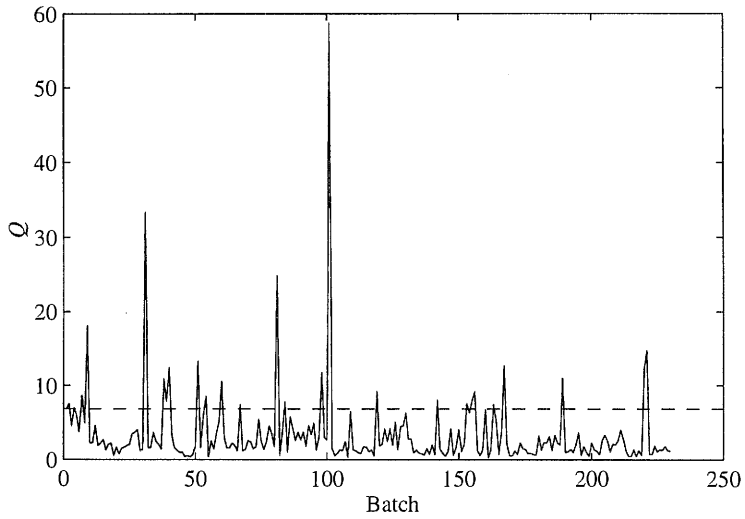

Fig. 3. Q (residual) control chart with 95% control limit.

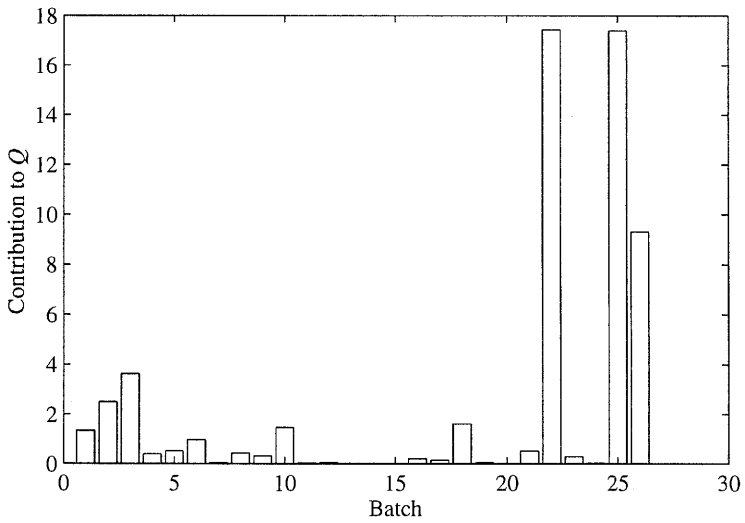

Fig. 4. Contributions to Q for batch 101.

3. Control charts focus on one variable across all batches. Contribution plots focus on all variables for one batch. Thus interpreting what problems exist in the data for a batch is easily done from contribution plots, but not easily done from control charts.
4. The use of multivariate statistics with contribution plots allows all of the variables collected to be used in an investigation. It allows the engineer to resist the

urge to keep the number of variables down to a manageable number (in our experience manageable means 5–10 variables). This is important for two reasons. First, variables are often left out of an analysis due to engineering judgment that they are not as important as other variables; however, if the process should unexpectedly change, it may become an important variable (and will usually be indicated by a large value of Q). We have seen several examples of this; for example, a pump had behaved well for a long period of time so there was very little variability in the flow rates based upon this pump. Removing the flow rates for this pump from the analysis would have caused us to miss the problems that occurred when the pump began to fail. A second reason one would not want to remove variables from the analysis is that often the collection of variables found via contribution plots may tell the story of the problem better than if we had used a subset of the variables and therefore found only a subset of the out-of-control variables. (Of course, one would not want to include in the data every possible variable; a method developed by Bopp and Grant (1989) provides guidance for which variables belong in the data set.)

6. SPC Using Scores

Up to now, we have used trend plots of T^2 and Q to identify out-of-control batches, followed by the use of contribution plots for cause identification. Plots of the Principal Components scores are also an effective tool for identifying special causes.

Scores are often plotted as a scatterplot of two dimensions. This provides a "two-dimensional window" that lets us observe the structure of our 27-dimensional data. In this scatterplot, we often see clusters or other features. Problems that are similar in the data will usually cluster together, regardless of their relationship in time sequence. This is another major advantage of the multivariate approach over standard control charting techniques, which display data in time sequence only.

For example, consider the group of five batches in a cluster shown in the scatterplot of the scores of dimension 1 and 3 in Fig. 5. These five batches are non-sequential in time sequence; they are batches 31, 142, 147, 220 and 221. We would like to know what process variable(s) caused these five batches to differ from the normal process operating conditions, which is represented by the cluster of batches at the origin. So far we have discussed contributions to the score of a single batch. We extend this idea now to contributions for multiple batches.

When we compute the contributions, we can replace x_{ij} with a suitably chosen average, or other linear combination of the data. In this case, we would use the average value $\bar{x}_{\cdot j}$, where the averaging is over the batches of interest. This in effect compares the average value of the batches of interest to the mean of all the process variables, which is zero for each (mean centered) process variable.

The average contribution to the scores in dimensions 1 and 3 for the five batches is shown in Fig. 6, and a time sequence plot of variable 13 is shown in Fig. 7, with the batches of interest indicated with asterisks. Variables 12 and 13 (both relating to silver pressure) were the major contributors in dimension 1, while variables 4 (average

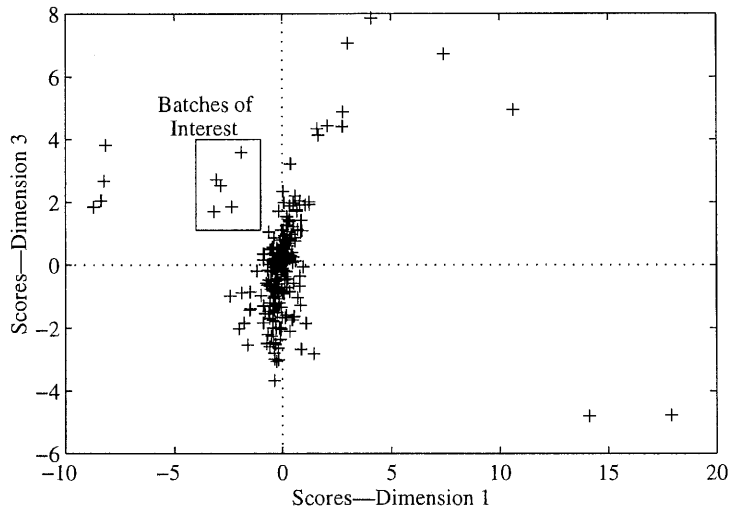

Fig. 5. Scatterplot of scores of dimensions 1 and 3.

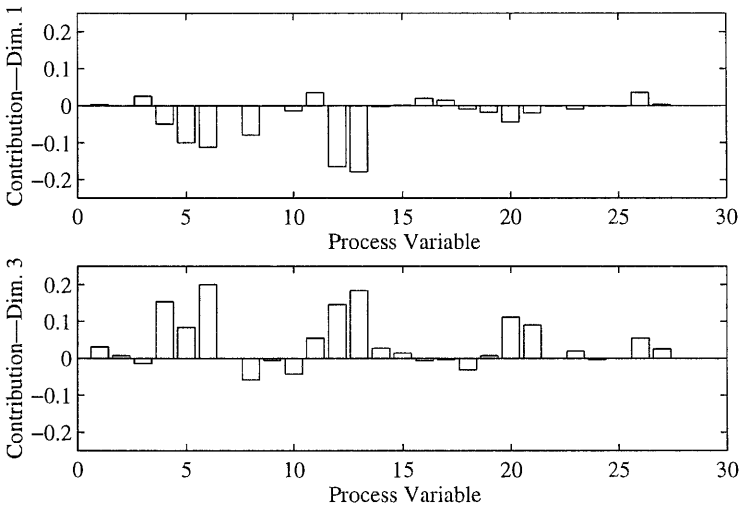

Fig. 6. Average contributions to dimensions 1 and 3 for cluster of 5 batches.

silver concentration during phase 1) and 6 (valve timing mismatch) also show up with large contributions in dimension 3. Unlike the problem with batch 208 discussed earlier, there is no indication that there was any problem in the salt delivery system. An engineering investigation discovered a "plug" in the silver delivery system, which caused pressures to back up, the incorrect amount of silver was delivered on time and the resulting chemical reaction did not take place properly. Each of the five batches had to be thrown out. By changing procedures, the likelihood of this happening again was minimized, thereby increasing the quality of the process and reducing waste.

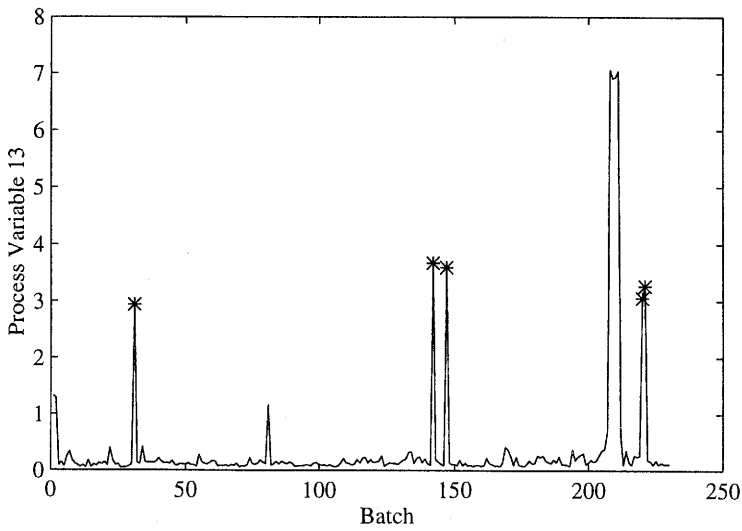

Fig. 7. Time sequence plot of process variable 13 with the 5 batches in the cluster indicated by asterisks.

We contend that without this scatterplot, it would have been extremely difficult, if not impossible, for an operator or engineer to recognize that the same pattern of problem variables occurred in these batches widely separated in time. Looking for non-sequential patterns in 27 different control charts is an *extremely* difficult perceptual task. With the scatterplot, the similarity of these batches is immediately obvious. With the contribution plot, the variables that are causing this out-of-control situation are easily detected.

The fact that there were five similar batches with this defect made the search for a solution more urgent, and allowed more resources to be directed at the problem. Had the engineers only known about the last batch (or in this case batches 220 and 221, which were sequential) the perceived magnitude of the problem, and hence the availability of the resources to solve the problem, would not have been so great.

The scores can be plotted in time sequence, with control chart limits if desired, to detect additional special causes in the data. Such a plot (without control chart limits) is shown in Fig. 8. Shewhart control chart rules for detection of special causes may be applied to this chart. Univariate EWMA or CUSUM techniques may be used on the scores as well to detect special causes. One such special cause that warrants further investigation is the level shift that occurs in score 3 at batch 74.

Score plots are very useful for detecting shifts and drifts in the process, while T^2 and Q are usually less useful for detecting shifts and drifts. (Note that for this example, prior to batch 74 the scores were running at approximately -2 and after the change they were at approximately $+2$, so T^2 would *not* show this process shift as being out-of-control.) As before, knowing that score 3 has shifted is not enough

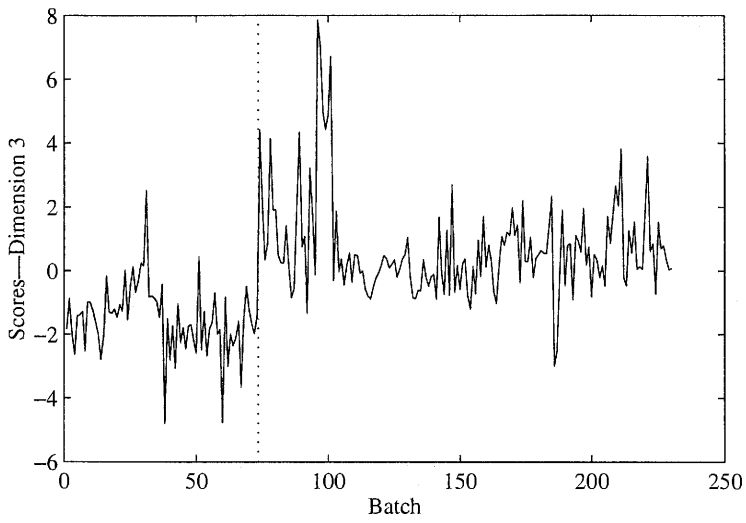

Fig. 8. Time sequence plot of score 3, with dotted line between batches 73 and 74.

information for the engineer to begin diagnosing the problem. Thus we need to tie the change in score 3 at batch 74 back to the original process variables using contributions.

In this situation, we want to compare the process before batch 74 to the process after batch 74, and so we would use the average x_{ij} before the shift minus the average x_{ij} after the shift. We may decide to choose 10 batches prior to the shift and 10 batches after the shift to compute the contributions. The number 10 is entirely based upon the judgment of the engineer and knowledge of the process; other sample sizes may be appropriate in other situations. In this case, the linear combination of the data has weights of $+0.1$ for batches 64 through 73, -0.1 for batches 74 through 83 and 0 elsewhere. The contributions are the k values of

$$\left(\sum_{i=64}^{73} x_{ij} - \sum_{i=74}^{83} x_{ij}\right) p_{j3} \Big/ 10, \quad j = 1 \ldots, k$$

and are shown in Fig. 9, in which it is clear that variables 10, 18 and 27 are contributing to the shift. A time sequence plot of variable 18 is shown in Fig. 10, confirming that it did indeed shift at batch 74. All three of the variables with the largest contributions related to the temperature control in the batch, and in fact the engineering investigation did turn up problems in this area which took a while to solve. Note that in Fig. 10, the process variable returns to its state before batch 74 somewhere around batch 105. Also around batch 105, the scores in dimension 3 do not return to their level before batch 74, indicating that other process phenomena are affecting the scores.

Sometimes, a drift upwards or downwards will be seen in the time sequence plot of the scores. In that situation, a linear combination of the data that estimates a slope will be useful for determining which process variables were drifting, causing the

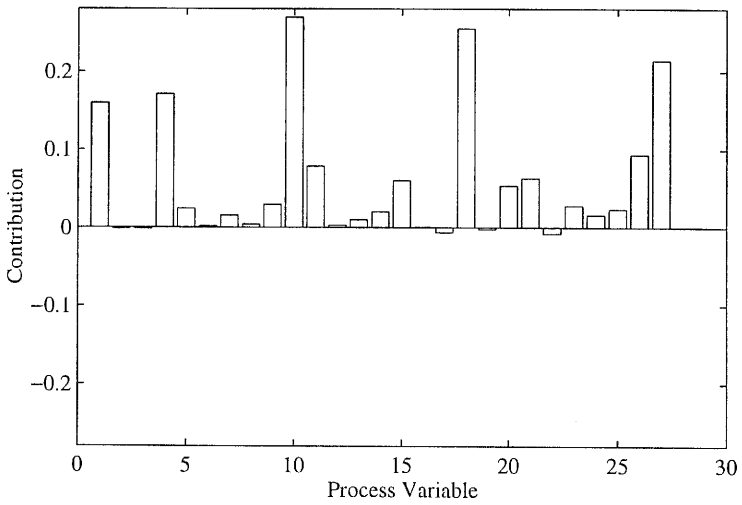

Fig. 9. Contributions to change in score 3 at batch 74.

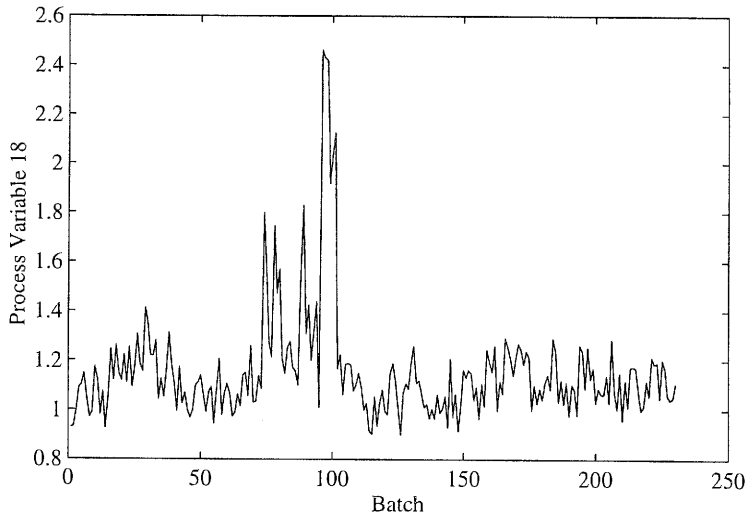

Fig. 10. Time sequence plot of process variable 18.

score to drift. Other linear combinations of the data can be used when appropriate. For example, if we saw n batches in a row drifting upwards or downwards, we could use the first order orthogonal polynomial for n data points as the weights for the linear combination of the observations.

7. Scaling of Contribution Plots

We have used two different scaling methods for plotting the contributions to the scores. The idea is to make the variables with the biggest contributions stand out visually. Both methods have been found to be useful, and both have been found to have certain drawbacks. The two methods either "zoom in" or "zoom out" on the plot, but leave the pattern of bar heights unchanged. The plots in this article use method 1 below. Q contribution plots are arbitrarily scaled.

Method 1—Maximum Contribution Scaling. For dimension d, we plot $x_{ij}p_{jd}/\max_{ij}|x_{ij}p_{jd}|$ for $j = 1, \ldots, k$. In this way, we compare the contributions for batch i to the maximum, in absolute value, of the contributions for all of the batches. If the contribution for batch i is ± 1, then this represents the worst deviation from the mean of all of the batches over all variables. The drawback to this method is that a batch which is a problem but not the biggest problem in the data set may appear to have only "small" bars. Another problem is that all dimensions will have the same scaling; we know that all dimensions are not equally important and the common scaling could be misleading.

Method 2—Within-Batch Scaling. For dimension d, we plot $x_{ij}p_{jd}/\sum_j |x_{ij}p_{jd}|$ for $j = 1, \ldots, k$. The biggest bars in this method are truly the ones which contribute most to the score for this particular batch and the height of the bar is roughly the proportion of the variable's contribution (it would be exactly the proportion if all of the values $x_{ij}p_{jd}$ had the same sign). However, the drawback to this approach is that a batch which really is just random noise on all variables will still show "big" bars.

8. Other Uses of Contributions

Although this article deals with Principal Components Analysis, contributions can be used with other dimension reduction techniques. Two that come to mind are Factor Analysis (FA) and Partial Least Squares (PLS). FA has been used for many years, mostly in the social sciences, while PLS is much newer and is slowly gaining attention. PLS tends to produce scores that contain more information about process variables that affect one or more response variables, such as quality characteristics. We refer the reader to Kresta *et al.* (1991) and Piovoso *et al.* (1992) for examples of using PLS for statistical process control. In addition, we believe that contribution plots might be helpful in any exploratory analysis of a multivariate data set, regardless of the application. Thus we can envision their use in analyzing data from marketing, econometrics, chemometrics, physical sciences, etc.

9. Comparison to Hayter and Tsui

Hayter and Tsui (1994) address the same issue as we do. They suggest a technique for identifying variables that are unusual once a multivariate statistic has indicated an out of control situation. There are two drawbacks to utilizing their method: it requires

a simulation, and it misses situations where there are subtle but critical shifts in the correlation structure of the variables. We pick these shifts up by signaling an out of control observation via the Q statistic, and then proceed in making a contribution plot for Q identifying the out-of-control variables in the out of control observation. We contend that out of control observations triggered by Q are just as important, and maybe more so, than out of control observations triggered on T^2. For example, when a pump begins to fail its output begins to change its correlation relative to its input. This is picked up very quickly by Q.

It is simple to illustrate the above point with the Linnerud Health Club Data given in Jackson (1991, p. 267). If one does a two-dimensional PCA analysis on the twenty autoscaled weight, pulse, and waist measurements, one can generate the Q Plot shown in Fig. 11. Observation #9 (i.e. the 176 pounder with the 31 inch waist) is clearly indicated as being "out of control". A variable Contribution Plot, Fig. 12, clearly shows that weight and waist are unusual. Autoscaled observation histories of these variables clearly show the reason for this out of control signal. The weight and the waist of this health club member move in opposite directions to what would have been expected. The maximum standardized value of the variables for this observation are all less than 2.0 in magnitude. This observation is clearly unusual and would not have been triggered with the Hayter and Tsui approach with a 0.05 error rate. In fact, the critical value for a 0.05 error rate in the Hayter and Tsui approach is about 2.3.

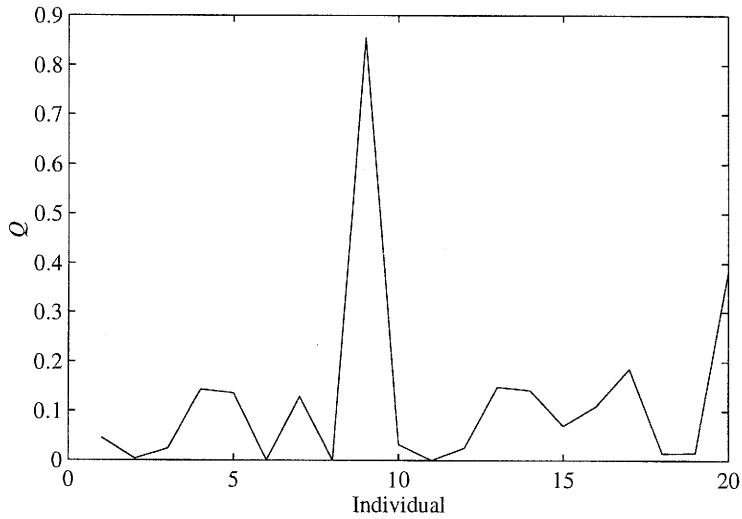

Fig. 11. Q Plot for Linnerud Health Club Data.

10. Summary and Recommendations

We have introduced a new statistic and plot to provide the missing link between multivariate out-of-control signals and the original variables which cause the multivariate

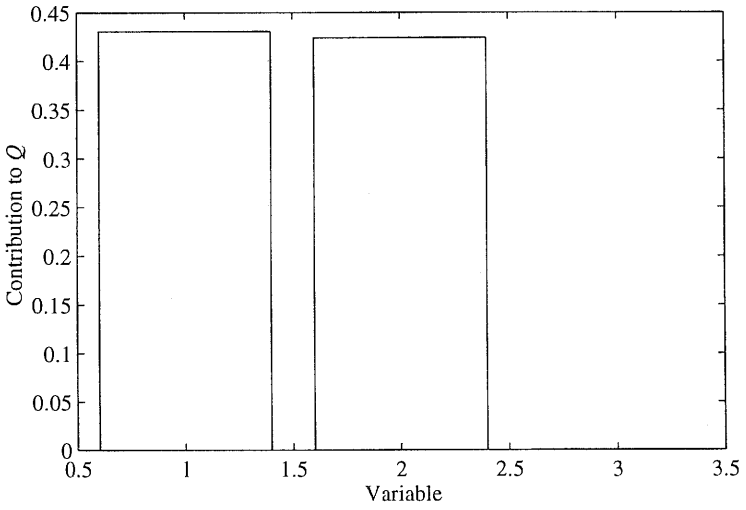

Fig. 12. Contribution to Q for observation 9 of Linnerud Health Club Data.

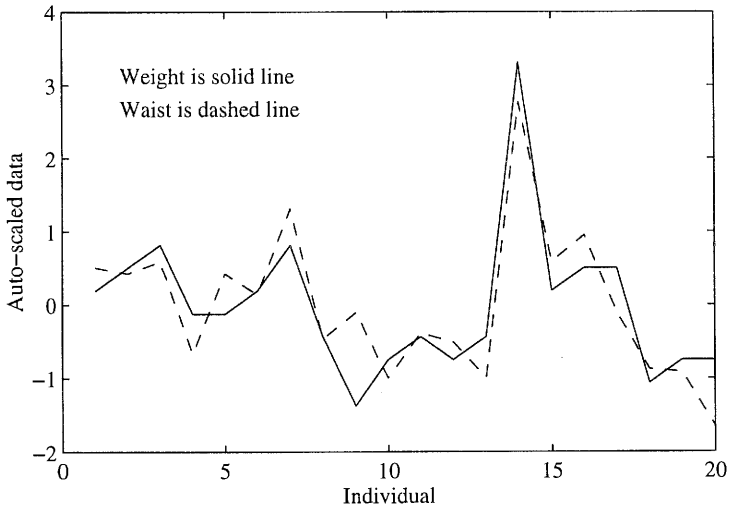

Fig. 13. Auto-scaled observation histories of Linnerud Health Club Data.

statistic to go out-of-control. We believe that this overcomes one of the primary difficulties in using multivariate statistics for multivariate quality control applications. Indeed, our experience is that operators and engineers can interpret and make use of the information contained in contribution plots and in the T^2, Q and score plots, even if they cannot completely explain or understand the method of calculation.

The use of multivariate statistics with contribution charts has numerous advantages over ordinary Shewhart or other univariate charting procedures. Some of these advantages are: reduction of the number of charts needed to monitor the process; ability to handle correlated variables and detect multivariate outliers; charts which show all variables for the batch of interest; and grouping of similar problems together even if they are non-sequential. In addition, the contribution plot idea detects and diagnoses more problems than the Hayter and Tsui approach or the Fuchs and Benjamini approach. The contribution method also requires only standard PCA software and standard graphics capabilities; no simulations or new types of graphics need to be programmed.

There is a need for research regarding type I error rates for non-normally distributed data. Also, methods need to be developed to detect what constitutes a significant bar on the contribution plots.

Although the use of contribution plots could potentially lead to a lot of plots being generated, we believe that the engineer should use the plots as a means of navigating through the data to the pieces of information that will guide his investigation. Graphical methods of presentation are essential if process experts are to be effectively involved. When implemented in a highly interactive graphical computing environment, our methods have proven to be a catalyst for linking engineers with their data, solving problems and increasing process knowledge.

References

Bopp A.L. and Grant R.P. (1989): *Statistical process control based on function analysis.* — TAPPI J., Vol.72, No.4, pp.77–79.

Crosier R.B. (1988): *Multivariate generalizations of cumulative sum quality control schemes.* — Technometrics, Vol.30, No.3, pp.291–303.

Devlin S.J., Gnanadesikan R. and Kettenring J.R. (1981): *Robust estimation of dispersion matrices and principal components.* — J. Amer. Stat. Assoc., Vol.76, No.374, pp.354–362.

Fuchs C. and Benjamini Y. (1994): *Multivariate profile charts for statistical process control.* — Technometrics, Vol.36, No.2, pp.182–195.

Hawkins D.M. (1993): *Regression adjustment for variables in multivariate quality control.* — J. Quality Technol., Vol.25, No.3, pp.170–182.

Hayter A.J. and Tsui K.L. (1994): *Identification and quantification in multivariate quality control problems.* — J. Quality Technol., Vol.26, No.3, pp.197–208.

Hopkins R.W., Miller P., Swanson R.E. and Scheible J.J. (1995): *Method of controlling a manufacturing process using multivariate analysis.* — United States Patent: 5, 442, 562.

Jackson J.E. (1991): *A User's Guide to Principal Components.* — New York: Wiley.

Kourti T. and MacGregor J.F. (1996): *Multivariate SPC methods for process and product monitoring.* — J. Quality Technol., Vol.28, No.4, pp.409–428.

Kresta J.V., MacGregor J.F. and Marlin T.E. (1991): *Multivariate statistical monitoring of process operating performance.* — Can. J. Chem. Eng., Vol.69, No.1, pp.35–47.

Lowry C.A., Woodall W.H., Champ C.W. and Rigdon S.E. (1992): *A multivariate exponentially weighted moving average control chart.* — Technometrics, Vol.34, No.1, pp.46–53.

MacGregor J.F. and Kourti T. (1995): *Statistical process control of multivariate process.* — Control Eng. Practice, Vol.3, No.3, pp.403–414.

Miller P. and Swanson R.E. (1993): *Contribution plots: The missing link in multivariate quality control.* — 37th Annual Fall Technical Conference, ASQC, Rochester, NY.

Piovoso M., Kosanovich K. and Yuk J. (1992): *Process data chemometrics.* — IEEE Trans. Instr. Meas., Vol.41, No.2, pp.262–268.

Wierda S.J. (1994): *Multivariate statistical process control—recent results and directions for future research.* — Statistica Neerlandica, Vol.48, No.2, pp.147–168.

Wise B.M., Ricker N.L., Veltkamp D.F. and Kowalski B.R. (1990): *A theoretical basis for the use of principal components models for monitoring multivariate processes.* — Proc. Contr. Quality, Vol.1, No.1, pp.41–51.

Wold S. (1978): *Cross-validatory estimation of the number of components in factor and principal components analysis.* — Technometrics, Vol.20, No.4, pp.397–405.

Woodall W.H. and Ncube M. (1985): *Multivariate CUSUM quality-control procedures.* — Technometrics, Vol.27, No.3, pp.285–292.