CONTRIBUTORS TO A SIGNAL FROM AN ARTIFICIAL CONTRAST

Jing Hu and George Runger
Arizona State University
Tempe, AZ

Eugene Tuv
Intel Corporation
Chandler, AZ

Keywords: Patterns, statistical process control, supervised learning, multivariate analysis.

Abstract: Data from a process or system is often monitored in order to detect unusual events and this task is required in many disciplines. A decision rule can be learned to detect anomalies from the normal operating environment when neither the normal operations nor the anomalies to be detected are pre-specified. This is accomplished through artificial data that transforms the problem to one of supervised learning. However, when a large collection of variables are monitored, not all react to the anomaly detected by the decision rule. It is important to interrogate a signal to determine the variables that are most relevant to or most contribute to the signal in order to improve and facilitate the actions to signal. Metrics are presented that can be used determine contributors to a signal developed through an artificial contrast that are conceptually simple. The metrics are shown to be related to traditional tools for normally distributed data and their efficacy is shown on simulated and actual data.

1 INTRODUCTION

Statistical process control (SPC) is used to detect changes from standard operating conditions. In multivariate SPC a $p \times 1$ observation vector $\mathbf{x}$ is obtained at each sample time. Some statistics, such as Hotelling's statistic (Hotelling, 1947), have been developed to detect whether the observation falls in or out of the control region representing standard operating conditions. This leads to two important comments. First, the control region is defined through an analytical expression which is based on the assumption of normal distribution of the data. Second, after a signal further analysis is needed to determine the variables that contribute to the signal.

Our research is an extension of the classical methods in terms of the above two points. The results in (Hwang et al., 2004) described the design of a control region based only on training data without a distributional assumption. An artificial contrast was developed to allow the control region to be learned through supervised learning techniques. This also allowed for control of the decision errors through appropriate parameter values. The second question is to identify variables that are most relevant to or most contribute to a particular signal. We refer to these variables as *contributors* to the signal. These are the variables that receive priority for corrective action. Many industries use an out-of-control action plan (OCAP) to react to a signal from a control chart. This research enhances and extends OCAP to incorporate learned control regions and large numbers of variables.

A physical event, such as a broken pump or a clogged pipe, might generate a signal from a control policy. However, not all variables might react to this physical event. Instead, when a large collection of variables are monitored, often only a few contribute to the signal from the control policy. For example, although a large collection of variables might be monitored, potentially only the pressure drop across a pump might be sensitive to a clogged pipe. The objective of this work is to identify these contributors in order to improve and facilitate corrective actions.

It has been a challenge for even normal-theory based methods to completely solve this problem. The key issue is the interrelationships between the variables. It is not sufficient to simply explore the marginal distribution of each variable. This is made clear

Hu J., Runger G. and Tuv E. (2005).
CONTRIBUTORS TO A SIGNAL FROM AN ARTIFICIAL CONTRAST.
In *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics*, pages 3-10
DOI: 10.5220/0001172900030010

in our illustrations that follow. Consequently, early work (Alt, 1985; Doganaksay et al., 1991) required improvement. Subsequent work under normal theory considered joint distributions of all subsets of variables (Mason et al., 1995; Chua and Montgomery, 1992; Murphy, 1987). However, this results in a combinatorial explosion of possible subsets for even a moderate number of variables. In (Rencher, 1993) and (Runger et al., 1996) an approach based on conditional distributions was used that resulted in feasible computations, again for normally distributed data. Only one metric was calculated for each variable. Furthermore, in (Runger et al., 1996) a number of reasonable geometric approaches were defined and these were shown to result in equivalent metrics. Still, one metric was computed for each variable. This idea is summarized briefly in a following section. Although there are cases where the feasible approaches used in (Rencher, 1993) and (Runger et al., 1996) are not sufficient, they are effective in many instances, and the results indicate when further analysis is needed. This is illustrated in a following section.

The method proposed here is a simple, computationally feasible approach that can be shown to generalize the normal-theory methods in (Rencher, 1993) and (Runger et al., 1996). Consequently, it has the advantage of equivalence of a traditional solution under traditional assumptions, yet provides a computationally and conceptually simple extension. In Section 2 a summary is provided of the use of an artificial contrast with supervised learning is to generate a control region. In Section 3 the metric used for contributions is presented. The following section present illustrative examples.

2 CONTROL REGION DESIGN

Modern data collection techniques facilitate the collection of in-control data. In practice, the joint distribution of the variables for the in-control data is unknown and rarely as well-behaved as a multivariate normal distribution. If specific deviations from standard operating conditions are not a priori specified, leaning the control region is a type of unsupervised learning problem. An elegant technique can be used to transform the unsupervised learning problem to a supervised one by using an artificial reference distribution proposed by (Hwang et al., 2004). This is summarized briefly as follows.

Suppose $f(x)$ is an unknown probability density function of in-control data, and $f_0(x)$ is a specified reference density function. Combine the original data set $x_1, x_2, ..., x_N$ sampled from $f_0(x)$ and a random sample of equal size N drawn from $f_0(x)$.If we assign $y = -1$ to each sample point drawn from $f(x)$ and $y = 1$ for those drawn from $f_0(x)$, then learning control region can be considered to define a solution to a two-class classification problem. Points whose predicted y are -1 are assigned to the control region, and classified into the "standard" or "on-target" class. Points with predicted y equal to 1 are are classified into the"off-target" class.

For a given point x, the expected value of y is

$$\begin{aligned}\mu(x) = E(y|x) &= p(y=1|x) - p(y=-1|x)\\ &= 2p(y=1|x) - 1\end{aligned}$$

Then, according to Bayes' Theorem,

$$\begin{aligned}p(y=-1|x) &= \frac{p(y=-1|x)}{p(x)}\\ &= \frac{p(x|-1)p(y=-1)}{p(x|-1)p(y=-1)+p(x|1)p(y=1)}\\ &= \frac{f(x)}{f(x)+f_0(x)} \qquad (1)\end{aligned}$$

where we assume $p(y = 1) = p(y = -1)$ for training data, which means in estimating $E(y|x)$ we use the same sample size for each class. Therefore, an estimate of the unknown density $f(x)$ is obtained as

$$\hat{f}(x) = \frac{1-\widehat{\mu}(x)}{1+\widehat{\mu}(x)} \times f_0(x), \qquad (2)$$

where $f_0(x)$ is the known reference probability density function of the random data and $\hat{\mu}(x)$is learned from the supervised algorithm. Also, the odds are

$$\frac{p(y=-1|x)}{p(y=1|x)} = \frac{f(x)}{f_0(x)} \qquad (3)$$

The assignment is determined by the value of $\hat{\mu}(x)$. A data x is assigned to the class with density $f(x)$ when

$$\widehat{\mu}(x) < v,$$

and the class with density $f_0(x)$ when

$$\widehat{\mu}(x) > v.$$

where v is a parameter that can used to adjust the error rates of the procedure.

Any supervised learner is a potential candidate to build the model. In our research, a Regularized Least Square Classifier (RLSC) (Cucker and Smale, 2001) is employed as the specific classifier. Squared error loss is used with a quadratic penalty term on the coefficients (from the standardization the intercept is zero). Radial basis functions are used at each observed point with common standard deviation. That is the mean of y is estimated from

$$\begin{aligned}\mu(x) &= \beta_0 + \sum_{j=1}^{n} \beta_j \exp\left(-\frac{1}{2}\|x-x_j\|^2/\sigma^2\right)\\ &= \beta_0 + \sum_{j=1}^{n} \beta_j K_\sigma(x, x_j) \qquad (4)\end{aligned}$$

Also, let $\beta = (\beta_1, \ldots, \beta_n)$. The β_j are estimated from the penalized least squares criterion

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{n} \beta_j \exp\left(-\frac{1}{2}\|x - x_j\|^2/\sigma^2\right)\right)^2 + \gamma\|\beta\|^2 \tag{5}$$

where n is the total number of observations in the training data set. If the x's and y are standardized to mean zero then it can be shown that $\widehat{\beta}_0 = 0$. Also, let the matrix K denote the $n \times n$ matrix with (i, j)th element equal to $K_\sigma(x_i, x_j)$. Then for a fixed σ the solution for β is

$$\widehat{\beta} = (K + n\gamma I)^{-1}\mathbf{y} \tag{6}$$

and this is used to estimate $\mu(x)$.

3 CONTRIBUTORS TO A SIGNAL

In this section, a metric is developed to identify variables that contribute to a signal from SPC based upon artificial contrasts. Suppose there are p correlated variables $(x_1, x_2, \ldots, x_p)$. Let x^* be an observed data point that results in a signal from the control scheme. Define the set

$$L_k = \{x | x_i = x_i^*, i \neq k\}$$

There are several reasonable metrics for the contribution of variable x_k to the out-of-control signal. We use

$$\eta_k(x^*) = \max_{x \in L_k} \frac{\widehat{f}(x)}{\widehat{f}(x^*)} \tag{7}$$

This measures the change from $\widehat{f}(x)/\widehat{f}(x^*)$ that can be obtained from only a change to x_k. If $\eta_k(x^*)$ is small then x_k^* is not unusual. If $\eta_k(x^*)$ is large, then a substantial change can result from a change to x_k and x_k is considered to be an important contributor to the signal.

From (2) it can be shown that $\widehat{\mu}(x)$ is a monotone function of the estimated density ratio $\widehat{f}(x)/\widehat{f}_0(x)$. Therefore, the value $x_k \in L_k$ that maximizes the estimated density ratio also maximizes $\widehat{\mu}(x)$ over this same set. In the special case that $f_0(x)$ is a uniform density the value of $x_k \in L_k$ that maximizes $\widehat{\mu}(x)$ also maximizes $\widehat{f}(x)$ over this set. Consequently, $\eta_k(x^*)$ considers the change in estimated density that can be obtained from a change to x_k.

From (3) we have that η_k is the maximum odds ratio obtained over L_k

$$\eta_k(x^*) = \max_{x \in L_k} \frac{\hat{p}(y = -1|x)/\hat{p}(y = 1|x)}{\hat{p}(y = -1|x^*)/\hat{p}(y = 1|x^*)} \tag{8}$$

To compare values of $\eta_k(x^*)$ over k the denominator in (8) can be ignored and the numerator is a monotone function of $\hat{p}(y = -1|x)$. Consequently, the value in L_k that maximizes $\eta_k(x^*)$ is the one that maximizes $\hat{p}(y = -1|x)$. Therefore, the $\eta_k(x^*)$ metric is similar to one that scores the change in estimated probability of an in-control point.

A point that is unusual simultaneously in more than one variable, but *not* in either variable individually, is not well identified by this metric. That is, if x^* is unusual in the joint distribution of $(x_1, \ldots, x_k)$ for $k \leq p$, but not in the conditional marginal distribution of $f(x_i | x_j = x_j^*, i \neq j)$ then the metric is not sensitive. This implies that the point is unusual in a marginal distribution of more than one variable. Consequently, one can consider a two-dimensional set

$$L_{jk} = \{x | x_i = x_i^*, i \neq j, k\}$$

and a new metric

$$\eta_{jk}(x^*) = \max_{x \in L_{jk}} \frac{\widehat{f}(x)}{\widehat{f}(x^*)} \tag{9}$$

to investigate such points. This two-dimensional metric would be applied if none of the one-dimensional metrics $\eta_k(x^*)$ are unusual. Similarly, higher-dimensional metrics can be defined and applied as needed. The two-dimensional metric $\eta_{jk}(x^*)$ would maximize the the estimated density over x_j and x_k. It might use a gradient-based method or others heuristics to conduct the search. The objective is only to determine the pair of variables that generate large changes in the estimated density. The exact value of the maximum density is not needed. This permits large step sizes to be used in the search space. However, the focus of the work here is to use the one-dimensional metrics $\eta_k(x^*)$'s. Because the contribution analysis is only applied to a point which generates a signal, no information for the set of one-dimensional η_k's implies that a two-dimensional (or higher) metric needs to be explored. However, the one-dimensional η_k's are effective in many cases, and they provide a starting point for all cases.

3.1 Comparison with a Multivariate Normal Distribution

In this section, we assume the variables follow a multidimensional normal distribution. Under these assumptions, we can determine the theoretical form of the metric $\eta_k(x^*)$. Given the estimate of the unknown density $\widehat{f}(x)$, define x_0 as

$$x_0 = \text{argmax}_{x \in L_k} \widehat{f}(x)$$

For a multivariate normal density with mean vector μ and covariance matrix Σ

$$x_0 = \text{argmin}_{x \in L_k} (x - \mu)' \Sigma^{-1} (x - \mu)$$

Therefore, x_0 is the point in L_k at which Hotelling's statistic is minimized. Consequently, x_0 is the same point used in (Runger et al., 1996) to define the contribution of variable x_k in the multivariate normal case. The use of the metric in (7) generalizes this previous result from a normal distribution to an arbitrary distribution.

4 ILLUSTRATIVE EXAMPLE

4.1 Learning the In-Control Boundary

To demonstrate that our method is an extension of the traditional method, first we assume that the in-control data follow a multivariate normal distribution. In the case of two variables, we capture a smooth, closed elliptical boundary. Figure (1) shows the boundary learned through an artificial contrast and a supervised learning method along with the boundary specified by Hotelling's statistic (Hotelling, 1947) for the in-control data.

The size of in-control training data is 400 and the size of uniform data is also 400. The in-control training data are generated from the two-dimensional normal distribution $\mathbf{X} = \mathbf{C} * \mathbf{Z}$ with covariance

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{C}\mathbf{C}' = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

and $\mathbf{Z}$ following two-dimensional joint standardized normal distribution with $\rho = 0$. The smoothing parameter for the classifier is $\gamma = 4/800$. The parameter for the kernel function is $\sigma = \sqrt{8}$. The out-of-control training data are generated from the reference distribution. There are four unusual points: A $(3, 0)$, B $(3, 1)$, C $(3, 2)$, and D $(3, 3)$.

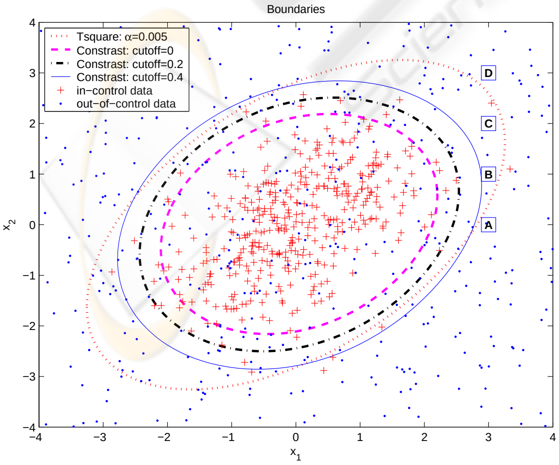

Figure 1: Learned Boundaries and Hotelling's Boundary

Table 1: Type I error for In-control Data

cut-off value	0	0.2	0.4
the training data	0.085	0.0325	0.015
the testing data	0.1	0.0525	0.025

Table 2: Type II error for Out-of-control Data with Different Shifted Means

cut-off value	0	0.2	0.4
(1,0)	0.785	0.895	0.96
(1,1)	0.7275	0.8325	0.8975
(2,0)	0.4875	0.6125	0.7325
(2,2)	0.3225	0.45	0.565
(3,0)	0.1025	0.215	0.325
(3,3)	0.055	0.1025	0.185

Testing data sets are used to evaluate performance, that is, Type I error and Type II error of the classifier. They are generated from similar multivariate normal distributions with or without shifted means. Each testing data set has a sample size of 400.

Table 1 gives the Type I error for the training data and for the testing data whose mean is not shifted. It shows that the Type I error decreases when the cut-off value of the boundary increases. Table 2 gives the Type II error for the testing data with shifted mean. It shows that for a given shift, the Type II error increases when the cut-off value of the boundary increases. It also illustrates that, for a given cut-off value, the further the mean shifts from the in-control mean, the lower the Type II error.

4.2 Contribution Evaluation

The probability density function of the in-control data $f(x)$ is estimated by (2). For the normal distribution in Section 3.1 examples are provided in the cases of two-dimensions (Figure 2)and 30-dimensions (Figure 3).

For the case of two dimensions, Figure (1) shows 4 points at $(3, 0)$, $(3, 1)$, $(3, 2)$, $(3, 3)$. The corresponding curves for $\widehat{f}(x)$ for each point are shown in Figure (4) through Figure (7). These figures show that the variable that would be considered to contribute to the signal for points $(3, 0)$ and $(3, 1)$ is identified by the corresponding curve. For point $(3, 2)$ the variable is not as clear and the curves are also ambiguous. For the point $(3, 3)$ both variables can be considered to the signal and this is indicated by the special case where all curve are similar. That is, no proper subset of variables is identified and this is an example where a higher-dimensional analysis (such as with $\eta_{jk}(x^*)$) is useful.

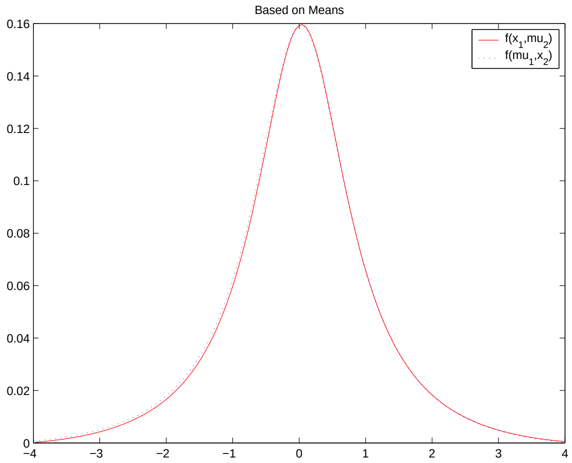

Figure 2: Density estimate for two dimensions

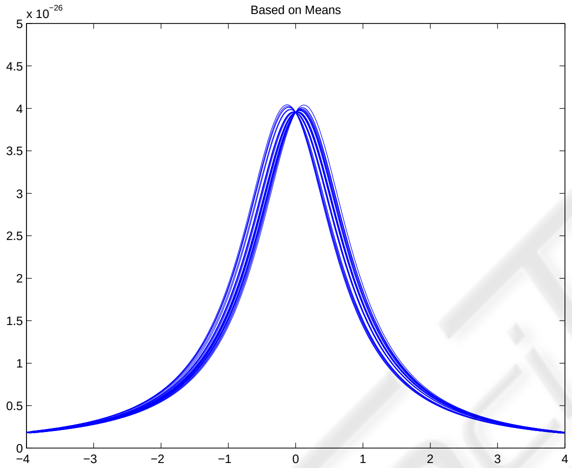

Figure 3: Density estimate for thirty dimensions

4.3 Example in 30 Dimensions

For a higher dimensional example, consider $p = 30$ dimensions. Out-of-control points are generated and density curves are produced for each variable. These curves are proportional to the conditional density with all the other variables at the observed values. For $p = 30$ the size of in-control training data is 200 and the size of the uniform data is also 200. Curves for out-of-control points

$$A = (3, 0, \dots, 0)$$

$$B = (3, -3, 0, \dots, 0)$$

$$C = (3, 3, \dots, 3)$$

are generated. For $p = 30$ dimensions the density curves are shown in Figure (8) through Figure (10).

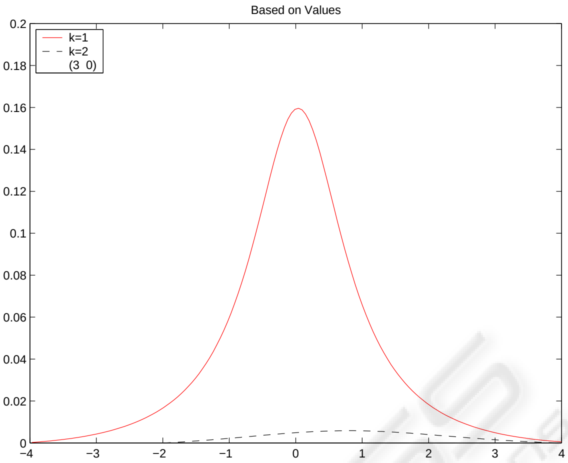

Figure 4: $f(x_1, 0)$ and $f(3, x_2)$

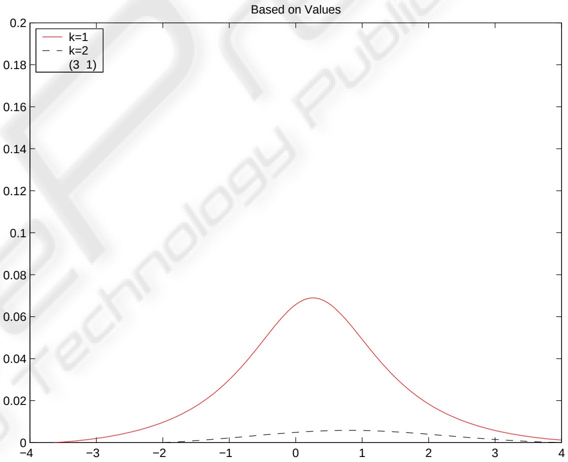

Figure 5: $f(x_1, 1)$ and $f(3, x_2)$

Note that the changes in density match the contributors to an unusual point. Note that for point C the density metric does not indicate any subset of variables as contributors. This is a special case and such a graph implies that all variables contribute to the signal from the chart because these graphs are only generated after a signal from a control has been generated. Such a special case is also distinguished from cases where only a proper subset of variables contribute to the signal.

For the particular case of $p = 30$ dimensions, values of $\eta_k(x_i)$ are calculated for these points and $k = 1, \dots, 30$ in Figure (11) through Figure (13). The results indicate the this metric can identify variables that contribute to the signal. For point C similar comments made for the density curves apply here. The metric does not indicate any subset of variables as contributors. This is a special case and such a graph

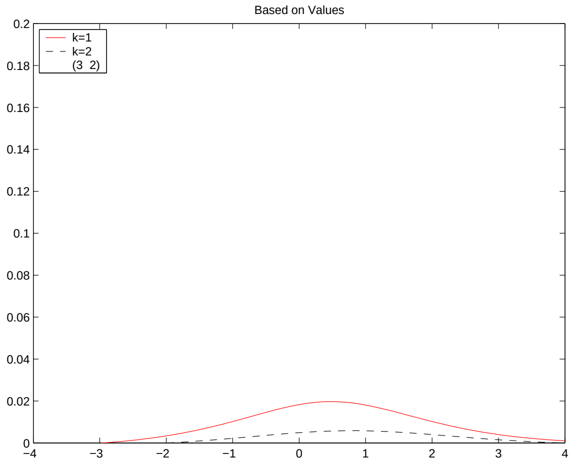

Figure 6: $f(x_1, 2)$ and $f(3, x_2)$

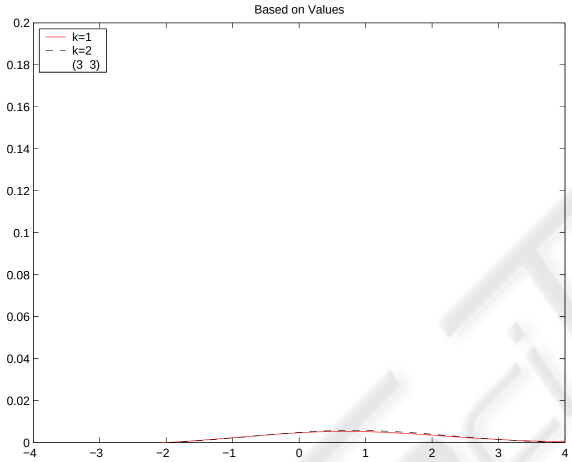

Figure 7: $f(x_1, 3)$ and $f(3, x_2)$

implies that all variables contribute to the signal from the chart.

5 MANUFACTURING EXAMPLE

The data set was from a real industrial process. There are 228 samples in total. To illustrate our problem, we use two variables. Here, Hotelling T^2 is employed to find out in-control data. The mean vector and covariance matrix are estimated from the whole data set and T^2 follows a χ^2 distribution with two degrees of freedom. The false alarm, α, is set as 0.05 in order to screen out unusual data. Figure (14) displays the Hotelling T^2 for each observation. From the results, we obtain 219 in-control data points that are used as the training data.

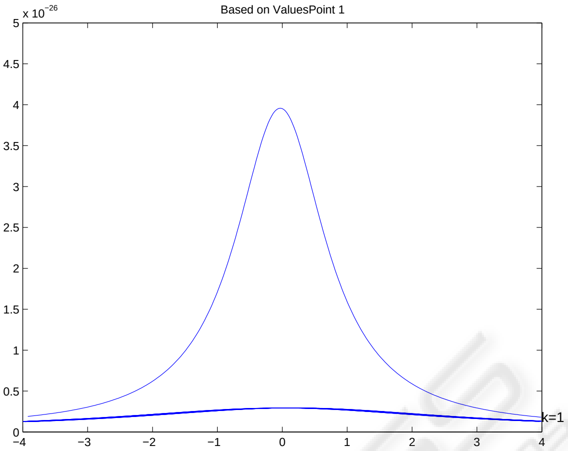

Figure 8: Density $f(x)$ as a function of x_k for $k = 1, 2, \ldots, 30$ for Point A

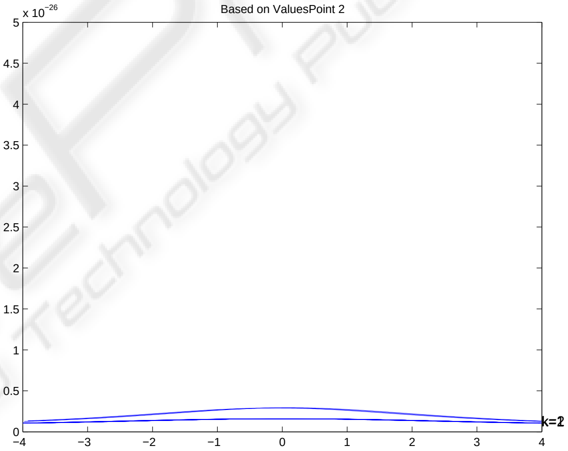

Figure 9: Density $f(x)$ as a function of x_k for $k = 1, 2, \ldots, 30$ for Point B

Figure (15) shows the learned boundaries with different cut-off values and the Hotelling T^2 boundary with α being 0.005. The learned boundary well captures the characteristic of the distribution of the in-control data. We select the learned boundary with cut-off $v = 0.4$ as the decision boundary and obtain three unusual points: Point 1, 2, and 4. The metric is applied to Point 2 and 4 and Table (3) and it demonstrates η values for each dimension for each point. Figure (16) and Figure (17) demonstrate $f(x_1, x_2)$ when as functions of x_1 and x_2 for Point 2 and 4, respectively. For Point 2, η_1 is significantly larger than η_2 so the first variable contributes to the out-of-control signal. For Point 4, η_1 and η_2 are close so both variables contributes to the out-of-control signal.

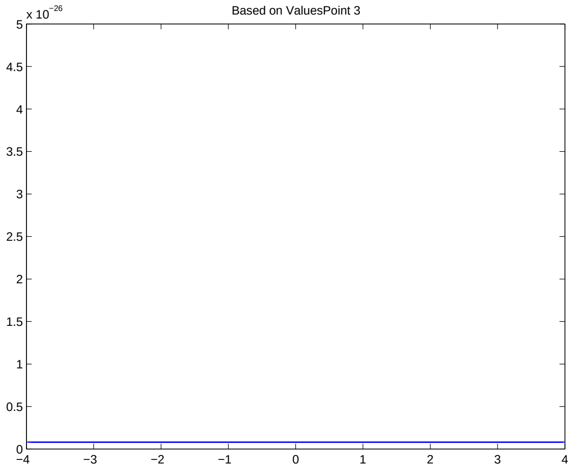

Figure 10: Density $f(x)$ as a function of x_k for $k = 1, 2, \ldots, 30$ for Point C

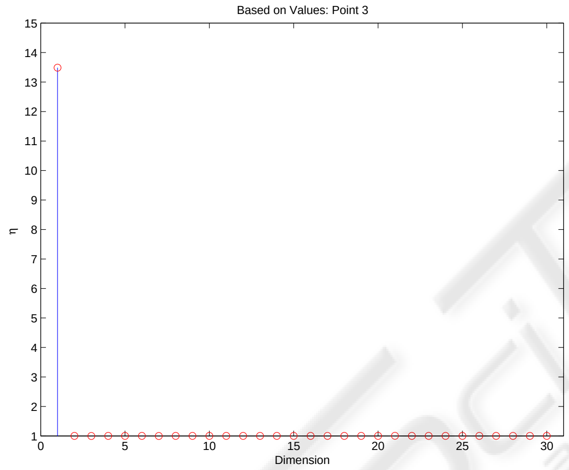

Figure 11: Contributor metric η_k for variables $k = 1, 2, \ldots, 30$ for Point A

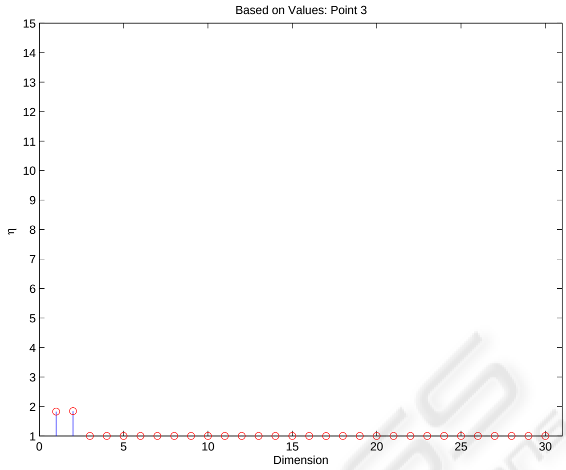

Figure 12: Contributor metric η_k for variables $k = 1, 2, \ldots, 30$ for Point B

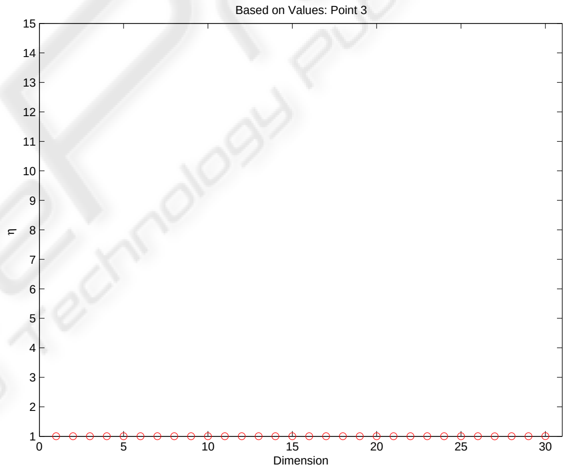

Figure 13: Contributor metric η_k for variables $k = 1, 2, \ldots, 30$ for Point C

6 CONCLUSION

A supervised method to learn normal operating conditions provides a general solution to monitor systems of many types in many disciplines. In addition to the decision rule it is important to be able to interrogate a signal to determine the variables that contribute to it. This facilitates an actionable response to a signal from decision rule used to monitor the process. In this paper, contributors to a multivariate SPC signal are identified from the same function that is learned to define the decision rule. The approach is computationally and conceptually simple. It was shown that the method generalizes a traditional approach for traditional multivariate normal theory. Examples show that the method effectively reproduces solutions for known cases, yet it generalizes to a broader class of problems. The one-dimensional metric used here would always be a starting point for such a contribution analysis. Future work is planned to extend the metric to two- and higher-dimensions to better diagnose contributors for cases in which the one-dimensional solution is not adequate.

Table 3: η for Point 2 and 4

	η_1	η_2
Point 2	16.791	1.0001
Point 4	3.6737	1.6549

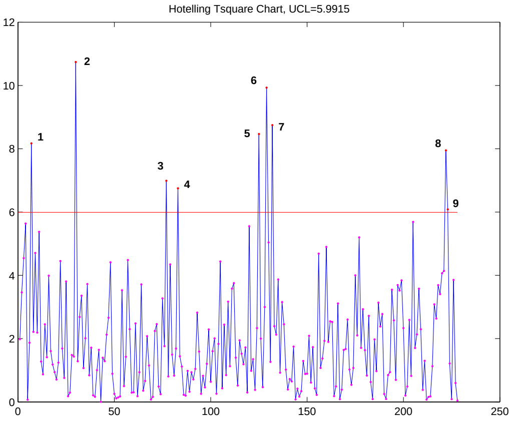

Figure 14: Hotelling T^2

REFERENCES

Alt, F. B. (1985). Multivariate quality control. In Kotz, S., Johnson, N. L., and Read, C. R., editors, *Encyclopedia of Statistical Sciences*, pages 110–122. John Wiley and Sons, New York.

Chua, M. and Montgomery, D. C. (1992). Investigation and characterization of a control scheme for multivariate quality control. *Quality and Reliability Engineering International*, 8:37–44.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of AMS*, 39:1–49.

Doganaksay, N., Faltin, F. W., and Tucker, W. T. (1991). Identification of out-of-control quality characteristics in a multivariate manufacturing environment. *Communications in Statistics-Theory and Methods*, 20:2775–2790.

Hotelling, H. (1947). Multivariate quality control-illustrated by the air testing of sample bombsights. In Eisenhart, C., Hastay, M., and Wallis, W., editors, *Techniques of Statistical Analysis*, pages 111–184. McGraw-Hill, New York.

Hwang, W., Runger, G., and Tuv, E. (2004). Multivariate statistical process control with artificial contrasts. under review.

Mason, R. L., Tracy, N. D., and Young, J. C. (1995). Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology*, 27:99–108.

Murphy, B. J. (1987). Selecting out-of-control variables with T^2 multivariate quality control procedures. *The Statistician*, 36:571–583.

Rencher (1993). The contribution of individual variables to hotelling's T^2, wilks' Λ, and R^2. *Biometrics*, 49:479–489.

Runger, G. C., Alt, F. B., and Montgomery, D. C. (1996). Contributors to a multivariate control chart signal. *Communications in Statistics - Theory and Methods*, 25:2203–2213.

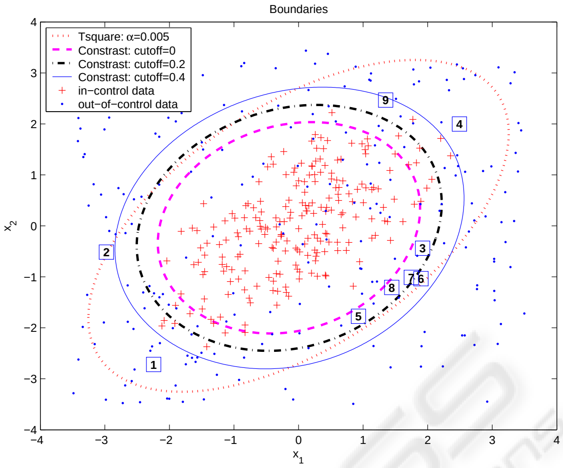

Figure 15: Learned Boundaries and Hotelling T^2

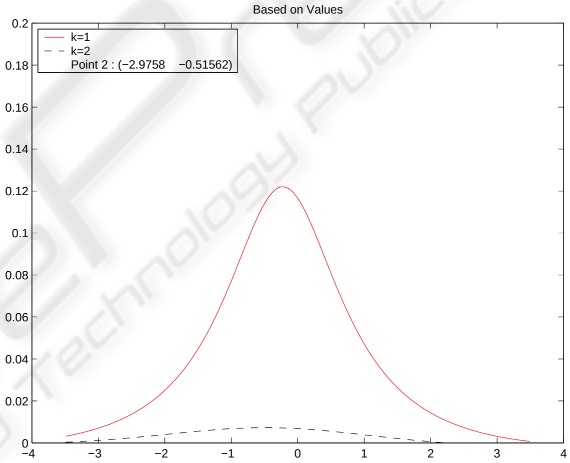

Figure 16: Density $f(x)$ as a function of x_1 and x_2 for Point 2

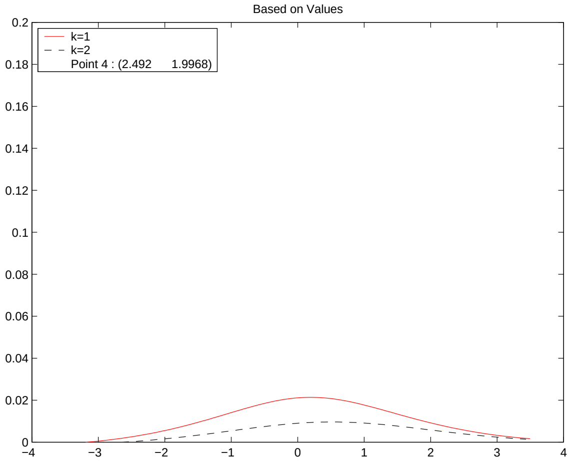

Figure 17: Density $f(x)$ as a function of x_1 and x_2 for Point 4