Application of Multivariate Statistics for Efficient Alarm Generation

Sandeep R. Kondaveeti ∗ Sirish L. Shah ∗,1 Iman Izadi ∗∗

∗ Department of Chemical and Materials Engineering, University of Alberta, Edmonton AB, Canada, T6G 2G6
∗∗ Matrikon Inc., Suite 1800, 10405 Jasper Avenue Edmonton, AB, Canada T5J 3N4

Abstract: Ever since the introduction of Distributed Control Systems (DCS), there has been little motivation for limiting the number of alarms that could be configured for process monitoring. As a result, plant operators are overwhelmed with many alarms during process upsets. Although a handful of these alarms are informative, many of them are a nuisance to the operator. However, all the alarms need to be acknowledged. Generally, check limits on univariate alarms are based on statistical quality control (three sigma limits, also known as Shewhart charts). While annunciating a univariate alarm on a particular variable, the information from other variables is often ignored. Modern day process plants have variables which are highly correlated. This correlation structure can be exploited in the efficient management of alarms. This work demonstrates the advantages of monitoring the PCA based $T^2$ and Q statistic over individual process variables. Monitoring these higher level statistics will not only reduce the false alarm and missed alarm rates but also reduces the detection latency which is one of the main drawbacks of monitoring a filtered variable. Two simulation examples are shown to illustrate the utility of the proposed method.

Keywords: Alarm systems, process monitoring, fault detection, ROC curves, multivariable systems, PCA.

1. INTRODUCTION

With increasing complexity in process industries, the study of abnormal event reduction has attracted a lot of attention in the recent past. Abnormal events in process plants lead to a variety of consequences ranging from a simple subsystem breakdown to as worse as loss of human life. On the economic front, an unplanned plant shutdown can wipe out all the benefits realized by advanced process control strategies. All abnormal events are initiated by a fault. Thus timely fault detection and diagnosis is a crucial step in reducing potential abnormal events.

1.1 Fault detection and process monitoring

The term fault is generally defined as a departure from an acceptable range of an observed variable or a calculated parameter associated with a process [Himmelblau, 1978]. The criterion for delineating a fault is a subjective task and is by no means straightforward if the stochastic aspects are taken into account. Thus, the definition of a fault depends on the characteristics of the process variables, their acceptable ranges, and the accuracy of the statistic used for classification of a potential fault.

There is a lot of literature on process fault detection ranging from analytical methods to artificial intelligence and statistical approaches [Venkatasubramanian et al., 2003]. From a modeling perspective, there are methods that require accurate process models, semi-quantitative models, or qualitative models. At the other end of the spectrum, there are methods that do not assume any form of model information and rely only on historical process data. Owing to their ease of use, better diagnosis capability and simplicity, the signal processing and statistics based approaches are widely embraced for process monitoring.

To ensure process specifications and safety, faults in the process need to be detected, diagnosed and eradicated. These tasks are associated with process monitoring [Russel et al., 2000]. Thus the objective of any process monitoring scheme would be to ensure that the plant operators are informed of faults in plant behavior in the form of alarms before they get worse and lead to a subsystem/system failure.

1.2 Alarms and alarm systems

In the control room terminology, an alarm is a notification that a fault/abnormal event has occurred and an operator must take action. Thus, alarm monitoring which is part of process monitoring is the plant operators’ duty. The purpose of an alarm is to protect personnel, equipment or the process from unsafe conditions, or to alert the operator when the process tends to drift towards low quality production.

An alarm would be annunciated if the variable being monitored crosses the check limits (High alarm limit and
low alarm limit). These check limits are called alarm limits (or control limits in the statistical quality control terminology). The alarm limits depend on the type of variable being monitored (ex: raw signal, error signal or filtered signal) and in some situations there can be just one limit (High or Low) instead of both or there can be more than two limits (high-high, high, low, low-low etc). The count of alarms on a variable should ideally be the same as the number of actions available at the operators disposal. An alarm system is a collection of alarms configured on a set of variables which is to be monitored by the control room operator. A well designed alarm system is expected to provide precise information regarding the health of the process. Thus the control room operator relies on the alarm system for monitoring the process.

Control and monitoring practice in process industries changed drastically after the advent of Distributed Control System (DCS) in the late 70s. The DCS introduced software alarms - alarms that are created or changed by configuring a setting in a computer, rather than requiring a hard-wired signal to a panel. As a result, more alarms could be configured at no extra cost. This resulted in sloppy alarm design on most of the variables leading to “nuisance alarms” - alarms that do not tell the operator anything he/she does not already know, or which do not require operator action. As a consequence of this oversight, alarm flooding occurs during an abnormal situation. Alarm flooding is a condition where alarms appear on the control panels at a rate faster than the operator can comprehend or respond to. Alarm flooding prevents the operator from determining the root cause of the process upset and therefore limits the scope for effective and quick emergency response.

The purpose of this work is to show the usefulness of multivariate statistics for process monitoring in terms of reduced false alarm rates and missed alarm rates.

2. MULTIVARIATE STATISTICAL PROCESS CONTROL

The main disadvantage of univariate monitoring schemes is that for a process with many variables, the correlation between the process variables is not considered while setting the individual alarm limits. Moreover, the difficulty in univariate process monitoring increases with the complexity of the process. Multivariate quality control (MQC) methods overcome this disadvantage by monitoring several variables simultaneously [MacGregor and Kourti, 1995]. A simple way to see the advantages of MQC is to superimpose univariate control charts on top of each other and create a graph of all the points of each control chart in an area of space.

Fig. 1 shows three plots. Top left and the bottom right are individual scatter plots of multivariate data composed of two variables \((p = 2)\), \(x\) and \(y\). The top right plot shows the graph of \(x\) v \(y\). The individual control limits for each variable’s respective univariate control chart are shown in the control rectangle. Though there is a faulty data point in the given data, the univariate control limits are not violated. If the strong correlation between \(x\) and \(y\) is exploited, the control limit on the process could be narrowed down and one such control limit shown as an ellipse in the top right plot of Fig. 1 would be able to flag the fault.

Consider a process with \(p\) variables. Assume that each of these \(p\) variables is monitored with a false alarm rate of \(\alpha\). When the process is in control (no fault), the joint probability of occurrence of at least one false alarm on one or more variables is:

\[
1 - (1 - \alpha)^p \approx p\alpha
\]

for small \(\alpha\). The value of this joint probability of false alarm rate increases as the number of monitored process variables increase.

3. PCA BASED STATISTICAL PROCESS MONITORING (SPM)

Principal Component Analysis (PCA) is a well known dimensionality reduction technique [Wold et al., 1987]. By choosing the appropriate number of Principal Components (PCs), a given data set can be projected on to two orthogonal subspaces, the principal component subspace (PCS) and the residual subspace (RS). PCS captures the normal variation of the data where as RS ideally captures only noise. Let \(x \in \mathbb{R}^m\) denote a sample vector of \(m\) sensors. Collecting the normal operating data for \(N\) samples, a data matrix \(X\) can be formed with each row representing a sample and each column representing a sensor. For correlation based PCA, the matrix \(X\) is scaled to zero mean and unit variance. The matrix \(X\) can be decomposed into a score matrix \(T\) and a loadings matrix \(P\) as

\[
X = TP^T = TP^T + \hat{T}\hat{P}^T = \hat{X} + \tilde{X}
\]

Given a sample vector \(x\), it can be projected on PCS and RS as

\[
\hat{x} = PP^Tx \quad \text{and} \quad \tilde{x} = (I - PP^T)x
\]

Two fault detection indices, \(T^2\) statistic and \(Q\) statistic are used as a measure of variability of a given sample in PCS and RS respectively. They are defined as follows:

\[
Q\text{ statistic} = \text{SPE} = \|\hat{x}\|^2 = \|(I - PP^T)x\|^2 \quad \text{and} \quad T^2 = x^T PA^{-1} P^Tx
\]

Fig. 1. Univariate versus multivariate process monitoring
Table 2: Confusion matrix in the alarm system terminology where $A$ is a diagonal matrix of eigen values corresponding to the eigen vectors in $P$.

A higher value for the $T^2$ statistic suggests that the sample is farther away from the normal operating region in the PCS. If the same sample has a small value for the Q statistic, it means that the correlation structure is not broken. Unless the fault magnitude is very high, the break in correlation does not increase the $T^2$ statistic significantly. The Q statistic is very sensitive to any break in correlation structure in the sample thus a low value of Q statistic indicates that there is no change in correlation structure. Thus the $T^2$ statistic and Q statistic play complementary roles [Qin, 2003]. Monitoring Q statistic is especially useful because it is sensitive to most common faults such as sensor failure, sensor bias, leaks in flows etc.

4. RECEIVER OPERATING CHARACTERISTICS (ROC) CURVES

The ROC curve is a graphical tool for visualizing a classifiers’ performance. ROC curves have long been used in fields like signal detection theory, medical diagnostic testing and machine learning to depict the tradeoff between hit rates and false alarm rates [Fawcett, 2006]. ROC curves are well appreciated in domains where there are skewed class distributions and unequal classification error costs. In this work, fault detection is dealt as a two class classification problem. The true classes are fault and no-fault, the corresponding hypothesized classes are alarm and no-alarm and the classifier is an alarm limit. Thus, given a sample of a signal (either raw or processed), it is compared with an appropriate threshold (classifier or alarm limit) and is mapped to one of the two classes, alarm (in case the threshold is violated) and no-alarm (otherwise). The confusion matrix is shown in Fig. 2.

Entries in the confusion matrix change as the discriminating threshold (alarm limit) is varied. Various measures can be defined for the classifiers performance. The most important ones are the false alarm rate and the missed alarm rate. They are defined as follows:

\[
\text{False Alarm Rate (FAR)} = 100 \times \frac{FA}{TA+FA} \%
\]
\[
\text{Missed Alarm Rate (MAR)} = 100 \times \frac{MA}{MAR+TA} \%
\]

Both FAR and MAR are problematic to the operator and the best classifier has to be chosen in such a way so as to minimize both of them.

Traditionally, ROC curves are plotted as FAR v 100-MAR. However, in this work, for better visualization and analysis, ROC curves are represented equivalently by plotting FAR v MAR. The best alarm limit usually corresponds to the point on the ROC curve closest to the origin (FAR = 0 %, MAR = 0%) when both false alarms and missed alarms are equally undesirable. The alarm rates corresponding to the best alarm limit are hereafter called as Minimum FAR (MFAR) and Minimum MAR (MMAR). Following is a simple illustration of an ROC curve and how MFAR and MMAR can be calculated.

Fig. 3 shows the ROC curve plotted as the alarm limit is varied to discriminate between normal operating data ($N(0,1)$) and faulty data ($N(2.5, 2.25)$) where $N(\mu, \sigma^2)$ represents normal distribution with mean $\mu$ and variance $\sigma^2$. The inset shows both the normal operating and faulty data. The design procedure involves selecting an appropriate alarm limit using the ROC curve. For example if the operator considers both false alarms and missed alarms to be equally undesirable, the best alarm limit corresponds to the point nearest to origin on the ROC curve. However, if the operator is willing to face more false alarms than missed alarms due to safety concern, an appropriate alarm limit can be determined using the ROC curve. Here, equal importance is given to both false alarms and missed alarms and the corresponding MFAR and MMAR point is determined as shown in Fig. 3.

5. EFFECTIVENESS OF MULTIVARIATE TECHNIQUES

5.1 Linear System

The following equations describe the dynamics of a linear system with 10 measured variables:

\[
x_5 = x_1 + x_2 \\
x_6 = x_2 + x_3 \\
x_7 = x_3 + x_4 \\
x_8 = x_1 + x_2 + x_4 \\
x_9 = x_2 + x_4 \\
x_{10} = x_1 + x_2 + x_3
\]
$x_1, x_2, x_3$ and $x_4$ are independent inputs with sufficient excitation. The excitation is added in the form of uniform random disturbance with magnitude of not more than 10% of their respective nominal values. $x_5, x_6, x_7, x_8, x_9$ and $x_{10}$ are linear combinations of the inputs and are the process outputs. $X_1, X_2, ..., X_{10}$ are process measurement vectors with added measurement noise. The measurement noise is normally distributed with zero mean and standard deviation of about 1% of the average of all the process variables at normal operation. Two types of faults are considered here. The first one is a sensor bias and the second is a measured disturbance (throughput change).

**Sensor bias in the linear system** The process is simulated for 10000 sampling instants. A sensor bias of magnitude 3% is introduced in the variable $x_8$ from sampling instant 5001. Since there are 4 independent variables in the process, a PCA model with 4 principal components (PCs) should be able to explain most of the variance in the normal operating data.

Fig. 4 shows the performance of three variables of interest for monitoring. They are the Q statistic using a PCA model with 4 PCs, faulty variable $X_8$ and filtered faulty variable $X_8^\ast$. Here, a third order moving average filter is used to obtain $X_8^\ast$ from $X_8$. It is to be noted that, unlike other two variables, filtering introduces detection delay in $X_8^\ast$. The magnitude of detection delay usually depends on filter type and distributions of normal and faulty operating data. It can be seen from Fig. 4 that the ROC curve for Q statistic is closer to origin (FAR = 0, MAR = 0) than that of $X_8$ and $X_8^\ast$.

Fig. 5 shows the MFAR and MMAR values for process variables, Q statistic and $T^2$ statistic. The PCA is done on the normal operating data after zero centering and unit variance scaling of all the variables. For the sensor bias fault, the correlation structure is broken in the faulty samples, thus the distribution of Q statistic is expected to change as we move from normal operation data to that of faulty. The accuracy of the Q statistic depends on how well the model is represented by the chosen PCs. It is evident that the MFAR and MMAR for the Q statistic drop significantly from the 4th PC (it is within 1% MFAR and 1% MMAR). For this linear process, a PCA model with 4 PCs should be adequate. If more PCs are considered, there is a good chance of including noise in the prediction and hence in the Q statistic. In this example, it appears that any PCA model with 4 to 8 PCs can be used to compute the Q statistic. However if the signal to noise ratio is low, it is recommended to use the exact number of latent variables (4 PCs in this case).

**Throughput change in the linear system** In this case, appropriate step disturbances are given in three of the input variables (about 7% in $x_1$, 4% in $x_3$ and 3% in $x_4$) starting from sampling instant 5001 until 10000. As there is be no break in correlation structure, the Q statistic for the PCA model with 4 PCs do not show any change. However, the $T^2$ statistic with 4 PCs gives lowest values for MFAR and MMAR (see Fig. 6). It is to be noted that the fault magnitude is larger here compared to that in the sensor bias case.

5.2 Tennessee Eastman Process (TEP)

Application of multivariate statistics for fault detection in TEP has been explored in great detail [Russel et al., 2000]. This present work is more in an alarm monitoring perspective and uses a different control structure to simulate custom faults.

The TE process is a simulation environment of a real chemical process with masked components and dynamics. The Tennessee Eastman (TE) Plant-wide Industrial Process Control Problem was proposed in the early 90s as a challenge test problem for a number of control related topics [Downs and Vogel, 1993]. As shown in Fig. 7, the
Variable number
Alarm rates for individual process variables
1 2 3 4 5 6 7 8 9 10 20 40 60 80 100
PCA model with around 27 PCs gives good results and as we move from 27 to 37 PCs, the Q statistic gives

Alarm rates for T square statistic
(11.2, 10.9)

Fig. 6. MFAR and MMAR for throughput change case in the linear process

Alarm rates for Q statistic
(5.6, 9.5)

Alarm rates for individual process variables
1 2 3 4 5 6 7 8 9 10 20 40 60 80 100
Variable number
Alarm rates for individual process variables
1 2 3 4 5 6 7 8 9 10 20 40 60 80 100
PC process includes following units: a two-phase reactor, a flash separator, reboiled stripper, recycle compressor and a condenser. There are four gaseous reactants (A, C, D, and E), two liquid products (G and H), a byproduct F and an inert B. In all there are 41 measured variables and 12 manipulated variables in the original TE process.

The open loop process as such is non-linear and highly unstable. In this work, the control structure proposed in [Ricker, 1996] is adopted to stabilize the plant and study its response under two faulty situations. The TE simulator with control structure in place can be downloaded from http://depts.washington.edu/control/LARRY/TE/download.html#Multiloop. Only 9 out of 12 manipulated variables are used in this control structure. Thus in all there are 50 variables in the process. Variables 42 (D feed), 45 (A&C feed), 48 (Stripper underflow valve) show significant change in their distribution. It is to be noted that all these variables are manipulated variables for the process loss. Here again, the bottom stem plot (for Q statistic) in Fig. 9 shows relatively low values for MMAR and MFAR.

Bias in the pressure sensor Reactor pressure (variable 7) is one of the most important variable in the TE process. According to the benchmark problem ([Downs and Vogel, 1993]), a 3.4% increase in the reactor pressure from normal operating range will lead to a process shutdown. A small sensor bias of magnitude 2 kPa is introduced in variable 7 after 36 hours of normal operation. The top part in Fig. 8 shows the MFAR and MMAR for all the process variables. Variable 7 shows no significant change from normal operation to the faulty operation. The bottom stem plot in Fig. 9 shows the MFAR and MMAR for pressure sensor bias case in the linear process.

Product leak The second fault considered is the stripper underflow (variable 17) leakage. The product flow is also a controlled variable thus we do not expect any change in its value from normal operation to the faulty operation. The top stem plot in Fig. 9 shows the MFAR and MMAR for all the 50 variables in the process. Variables 42 (D feed), 45 (A&C feed), 48 (Stripper underflow valve) show significant change in their distribution. It is to be noted that all these variables are manipulated variables for the process and they increase in order to compensate for the product loss. Here again, the bottom stem plot (for Q statistic) in Fig. 9 shows relatively low values for MMAR and MFAR. PCA model with around 27 PCs gives good results and as we move from 27 to 37 PCs, the Q statistic gives
MFAR and MMAR values within the 1% level which is much better compared to monitoring any of the individual process variables.

6. CONCLUSIONS

As complexity in process plants increase, univariate process monitoring becomes a herculean task. It is necessary to switch to multivariate alarms which appear to take advantage of analytical redundancy in the process networks. ROC curves are useful for visualizing the performance of alarm limits and are particularly appreciated in cases with skewed class distributions and unequal classification error costs. The PCA based $T^2$ and $Q$ statistic incorporate information from several variables simultaneously to flag a fault and thus are more efficient compared to univariate alarming techniques. The two simulation case studies, i) a linear process and ii) the TE benchmark process illustrate the efficiency of PCA based $T^2$ and $Q$ statistic in minimizing nuisance alarms. Monitoring filtered variables is another way to reduce the false alarms and missed alarms but filtering introduces detection delay. Once a a multivariate alarm is annunciated, the contribution plots [Qin, 2003] can be used to pinpoint the variables responsible for the fault. As part of future work, alarm rationalization is to be done on a real process using alarm event and process data.

REFERENCES


