Multivariate Analysis of Process Data using Robust Statistical Analysis and Variable Selection

Leo H. Chiang, Randy J. Pell, and Mary Beth Seasholtz

The Dow Chemical Company
Analytical Sciences Laboratory
Corporate R&D
1897 Building
Midland, MI 48667
U.S.A.

Abstract: Historical plant data are useful in developing multivariate statistical models for on-line process monitoring, soft sensors, and process troubleshooting. For the first two purposes, historical data are used to build a model to capture the normal characteristics of the process. However, the presence of outliers can adversely affect the model. Various robust statistical techniques are investigated in this paper for outlier identification. For process troubleshooting and fault identification, it is crucial to identify the key process variables that are associated with the root causes. Genetic algorithms (GA) are incorporated with Fisher discriminant analysis (FDA) for this purpose. These techniques have been successfully applied at The Dow Chemical Company.

Keywords: Fault identification, robust estimation, genetic algorithms, data processing, bad data identification

1. INTRODUCTION

Process data are rapidly collected and stored for the chemical industry. These historical data are highly useful in developing multivariate statistical models such as principal component analysis (PCA) or partial least squares (PLS) for on-line process monitoring. One important step in applying these techniques is to extract the normal data for the off-line model building phase. Historical databases contain data from normal operating conditions, faulty conditions, various operating modes, startup periods, and shutdown periods. The presence of outliers further complicates the task of identifying the normal data. Outliers can disrupt the correlation structure of the PCA or PLS model and the result will be a model that does not accurately represent the process. To extract representative normal data, several outlier detection algorithms such as resampling by half-means (RHM), smallest half volume (SHV), and ellipsoidal multivariate trimming (MVT) can be used. A multiple outlier detection algorithm, closest distance to center (CDC), is proposed in this paper. CDC is conceptually similar to SHV but computationally more efficient than SHV. The use of the Mahalanobis distance in the initial step of MVT is known to be ineffective for detecting outliers. To overcome this limitation, CDC is incorporated with MVT. To increase the sensitivity for outlier detection for SHV, CDC, and MVT, a new modified scaling approach is proposed.

With the representative normal data identified and a model for the process constructed , the next step is to apply the model for on-line process monitoring. Once a fault is detected on-line, the immediate step is to determine the root cause. The objective of fault identification is to determine the variables that are most relevant to diagnosing the fault, thereby focusing the plant operators and engineers on the subsystem(s) where the fault has most likely occurred.

The contribution chart is a commonly used technique for fault identification. Previous results show that contribution charts perform well for simple faults, but are less effective for identifying complex process faults (MacGregor and Kourti, 1995). This demonstrates the need to look for an alternative method for identifying process faults. In this paper, GAs are incorporated with Fisher discriminant analysis (FDA) for process fault identification

2. METHODS

2.1 Effect of Scaling

Auto scaling is commonly applied to multivariate data. For a data sequence $\{x_i\}$, the auto scaling procedure follows:

$$d_i = \frac{x_i - m_x}{s}$$

where m_x is the mean of the variable and s is the standard deviation. For data that follow a normal distribution, the probability that $|d_i| > 3$ is about 0.27%. In the commonly used "3σ edit rule", an observation x is regarded as an outlier when $|d_i| > 3$. In the presence of multiple outliers, the 3σ edit rule can perform poorly. This is demonstrated in Fig. 1a, in which observations 1-960 are normal data and observations 961 to 1440 are outliers. By definition outliers are data that are not consistent with the majority of the data. The mean and standard deviation of the normal data are 41.1 and 0.55, respectively. With multiple outliers occurring on the same side of the mean, the estimate of the mean of the entire data sequence is increased to 42.3. These outliers also inflate the standard deviation estimate more than threefold to 1.87. The 3σ edit rule fails to detect the outliers (*i.e.*, $|d_i| < 3$ for all observations in Fig. 1b).

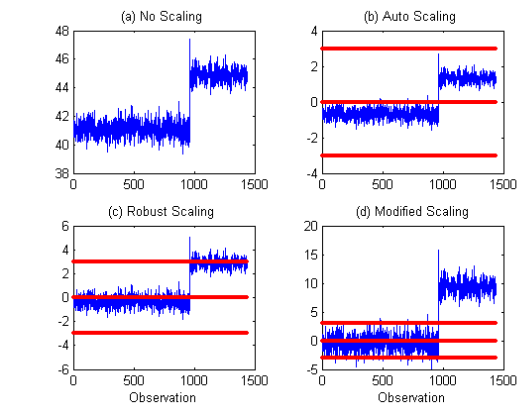

Fig. 1. Comparison of various scalings on the same variable. Observations 1-960 represent normal data and observations 961-1440 represent outliers. The solid lines represent the ±3σ thresholds.

To reduce the effect of multiple outliers, robust scaling has been suggested (Huber, 1989). In robust scaling, the mean is replaced with median and the standard deviation is replaced with median absolute deviation from the median (MAD):

$$s_{MAD} = 1.4826 \underset{i}{median}\left\{\left|x_i - x_{median}\right|\right\}$$

where x_{median} is the median of x. For the data used in Fig. 1, the median is 41.9, which is a fairly accurate location estimate for the normal data. The MAD for the data sequence is 1.17, which is a twofold overestimate. The 3σ edit rule with robust scaling, commonly referred to as the Hampel identifier (Pearson, 2001) fails to detect 60% of the outliers (see Fig. 1c).

Modfied scaling. To further increase the sensitivity in detecting outliers, a modified scaling is proposed here. For a variable with n observations, the $n/2$ observations that are nearest to the median are determined. The mean and standard deviation of these observations are used to autoscale the entire data sequence. For the data used in Fig. 1, the estimates of the mean and standard deviation are 41.3 and 0.39, respectively, which are close to the mean, 41.1, and standard deviation 0.55, found using the normal data only. With modified scaling, almost all the normal data are inside the 3σ thresholds while all the outliers are outside the 3σ thresholds (see Fig. 1d).

2.2 Robust outlier detection algorithms

Resampling by Half-Means (RHM): Given a data set of n observations and m process variables, a n by m matrix X is constructed. To start RHM, the first sample (i = 1) is obtained by randomly selecting half of the total observations. The sample i is written as a $n/2$ by m matrix $X_{sam}(i)$ and the mean $m(i)$ and standard deviation $s(i)$ vectors of the columns of $X_{sam}(i)$ are determined. The original data matrix X is autoscaled using $m(i)$ and $s(i)$, which results in a n by m autoscaled matrix X(i). The Euclidean distance is determined for each observation and a n by 1 vector of vector lengths $l(i)$ is obtained. The data are resampled for at least 2n times (Egan and Morgan, 1998). All the vector lengths are then stacked into an n by 2n matrix L. With sufficient resamplings, it is expected that the outliers will dominate in the upper (1-c) portion of L. For robust RHM, $m(i)$ and $s(i)$ are replaced with median and MAD, respectively.

Smallest Half Volume (SHV): In SHV (Egan and Morgan, 1998), the matrix X is first autoscaled and the Euclidean distance between each pair of observations i and j is determined. An n by n distance matrix D is formed and each column is sorted in ascending order. The column with the smallest sum for the first $n/2$ smallest distances is determined. These are the $n/2$ observations that are closest to each other in the multivariate space, which represent the most consistent portion of the normal data for most cases. In robust SHV and modified SHV, robust scaling and modified scaling are applied, respectively, to the matrix X first. The remaining steps are the same as the standard SHV.

Closest Distance to Center: CDC identifies the most consistent observations by calculating the distance of each observation from the center (i.e., mean for autoscaling and median for robust scaling) (Chiang *et al.*, 2003). In CDC, the matrix X is first autoscaled and the distance is determined for each observation. To equally weight the contribution for each variable to the distance, Euclidean distance (2-norm distance) can be used for each observation. This implementation is referred to as CDC_2. To emphasize

the most significant contribution of the variable to the distance, the maximum norm distance can be used for each observation. This implementation is referred to as CDC_m. The n/2 observations with the smallest distances represent the portion of the data that are closest to the center. Assuming that outliers are extreme observations that are far away from the majority of the data, these n/2 observations represent a portion of the normal data. Recall that the mean is not an accurate representation of the center of the data. A better implementation of CDC_2 and CDC_m is to use robust scaling or modified scaling prior to the distance determination steps.

Ellipsoidal Multivariate Trimming (MVT): MVT is an iterative procedure for the determination of a robust covariance matrix (Walczak and. Massart, 1995). In the first step of MVT, the Mahalanobis distance is determined for observation **x**

$$d_{mah} = (\mathbf{x} - \mathbf{x}^*)^T S^{*-1} (\mathbf{x} - \mathbf{x}^*)$$

where **x*** is the mean and S* is the covariance, calculated using all *n* observations. The *n*/2 observations with the smallest Mahalanobis distances are determined. Such observations are used to determine the new mean **x*** and new covariance S*. The Mahalanobis distance is recalculated using the new mean **x***, the new covariance S*, and the old **x**. The iterative procedure continues until *x** and S* stabilize.

With the presence of multiple outliers in the original data set, the covariance structure is disrupted. The use of Mahalanbis distance in the initial step of MVT can result in masking and swamping. As such, it is possible that further iterations in MVT do not improve the outlier detection proficiency. To overcome this weakness, robust outlier detection techniques such as RHM, SHV, CDC_2, or CDC_m can be used to determine the most consistent *n*/2 observations. These observations are then used to calculate **x*** and S*, upon which the initial Mahalanobis distance is calculated. In this paper CDC_m is used in conjunction with MVT. This implementation is referred to as CDC_m/MVT. Robust scaling and modified scaling are also applied in MVT and CDC_m/MVT.

2.3 Fault detection and fault identification

Principal Component Analysis: PCA is a well-known multivariate technique and detailed descriptions on the subject are available elsewhere (Chiang *et al.*, 2001; Beebe *et al.*, 1998). Only a brief review is given here. The PCA model is calculated using the singular value decomposition (SVD) on the autoscaled data matrix X

$$\frac{1}{\sqrt{n-1}} X = U\Sigma V^T$$

The loading vectors V corresponding to the *a* largest singular values are typically retained. These vectors are then stacked into an *m* by *a* loading matrix P. For on-line fault detection using the score space, the T^2 statistic can be calculated directly from the PCA representation (Jackson, 1959).

$$T^2 = \mathrm{x}^T P \Sigma_a^{-2} P^T \mathrm{x} = \mathrm{t}^T \Sigma_a^{-2} \mathrm{t}$$

where **t** is an *n* by 1 score vector, and S_a contains the first *a* rows and columns of Σ.

The portion of the observation space corresponding to the *m-a* smallest singular values can be monitored using the Q statistic (Jackson and Mudhlkar, 1979)

$$Q = \mathbf{x}^T (I - PP^T)\mathbf{x}$$

Contribution Chart: After a fault is detected (T^2 or Q statistics are larger than the threshold), the next step is to determine the root cause of the fault. While decentralized PCA techniques can often effectively isolate the location of the fault for large-scale systems (Georgakis *et al.*, 1996; Wachs and Lewin, 1999), the aim of the contribution chart is to determine the abnormal variables by calculating the contribution of each variable to the T^2 and Q statistics (Miller and Swanson, 1998). Detailed procedure to implement contribution charts is available elsewhere (MacMregor and Kourti, 1995; Chiang *et al.*, 2001).

Fisher Discriminant Analysis: FDA is a linear dimensionality reduction technique, optimal in terms of maximizing the separation between several classes (Duda and Hart, 1973). The FDA vectors are equal to the eigenvectors, $\mathbf{w_i}$, of the generalized eigenvalue problem

$$S_b \mathbf{w_k} = l_j S_w \mathbf{w_k}$$

where S_b is the between-class scatter matrix, S_w is the within-class scatter matrix, and the eigenvalues $ë_k$ indicate the degree of overall separability among the classes by projecting the data onto $\mathbf{w_k}$.

For classification, the discriminant function is calculated for class $j = 1$ to *c*. An observation **x** is assigned to class *j* that maximizes the discriminant function. Akaike's information criterion has been developed for automatically selecting the rank for FDA using the fitness function (Chiang, *et al.*, 2001)

$$f_{FDA} = s(a) - \frac{a}{n_{avg}}$$

where *s(a)* is the cross validation classification success rate at FDA rank *a* and n_{avg} is the average number of observations per class. The fitness function is calculated for $a = 1$ to $\min(m,n)$. The maximum fitness value, $f_{FDA,opt}$, represents the classification results at the optimal rank.

Genetic Algorithms: Once a fault is detected on line using PCA, GA/FDA can be used to determine the variables responsible for the root cause. A detailed review of GAs is available elsewhere (Leardi, 2001; Leardi *et al.*, 1992), only a brief review is given here. Two classes of data are used in FDA. Class 1 contains the training data representing the normal operating conditions. Class 2 contains data from a

time in which a fault is known or suspected to have occurred. GA/FDA begins with a first run by randomly creating n_p chromosomes. Only a subset of the original variables is selected in each chromosome. The performance of each chromosome is evaluated using a leave-1/5-out cross validation scheme with FDA. The fitness function $f_{FDA,opt}$ is then calculated for all chromosomes. Cross-over and mutations are performed over the evolutions in order to improve the chromosomes (*i.e.*, increase the fitness function $f_{FDA,opt}$). At the end of n_e evolutions, the chromosome with the highest $f_{FDA,opt}$ is saved.

The procedure is repeated for a second run. The final chromosome with the highest $f_{FDA,opt}$ at the end of n_e evolutions is saved. At the end of the n_r runs, n_r chromosomes are retained. A bar chart of the frequency of selection of each variable is then constructed. The plot represents the importance of each variable for distinguishing between the two classes. If the fitness function is high (*i.e.*, high success rate in cross-validated classification), these variables are often correlated with the root cause of the process fault. The variables are sorted according to the frequency of selection. The number of variables required to explain the root cause can be determined by maximizing the fitness function

$$f_{GA/FDA} = f_{FDA,opt}(m_{sub}) - \frac{m_{sub}}{n_{avg}}$$

where m_{sub} is the number of retained variables, corresponding to the first m_{sub} highest selected variables.

3. APPLICATIONS

Fig. 2 is a process flowsheet for the Tennessee Eastman Process (TEP). The TEP is based on an industrial process where the components, kinetics, and operating conditions were disguised for proprietary reasons (Downs and Vogel, 1993). The gaseous reactants A, C, D, and E and the inert B are fed to the reactor where the liquid products G and H are formed. The plant-wide control structure recommended in Lyman and Georgakis (1995) was implemented to generate the closed loop simulated process data for each fault.

TEP can simulate 21 process faults; Fault 6 is studied in detail in this paper. For Fault 6, there is a feed loss of reactant A in Stream 1 at t = 24 hr (see variable 1 in Fig. 3), which causes the control loop on Stream 1 to fully open the A feed valve (see variable 44 in Fig. 3). Because there is no reactant A in the feed, the reaction will eventually stop. This causes the gaseous reactants D and E to build up in the reactor, and hence the reactor pressure increases (see variable 7 in Fig. 3). The reactor pressure continues to increase until it reaches the safety limit of 2950 kPa, at this point the valve for Control Loop 6 is fully open. Clearly, it is very important to detect this fault promptly before the fault upsets the whole process. The proficiencies of contribution charts and GA/FDA are evaluated in terms of correctly identifying the root cause for Fault 6.

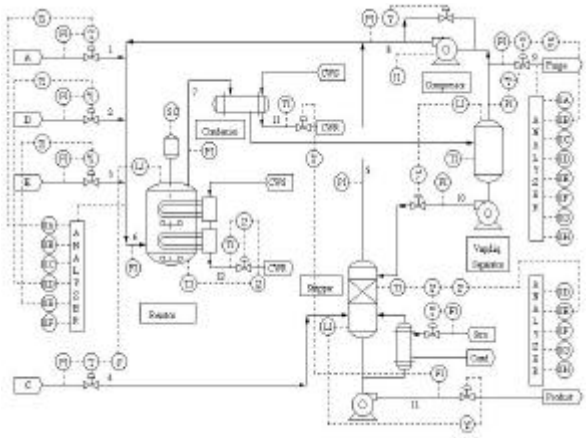

Fig. 2. A process flowsheet for the TEP

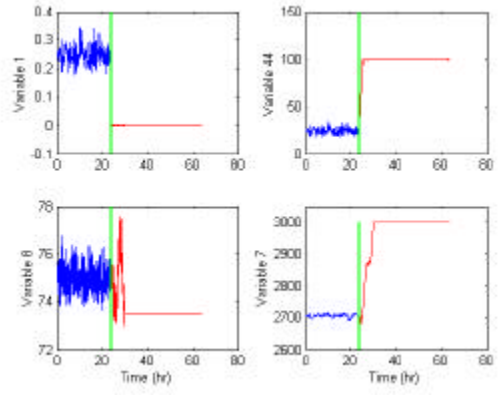

Fig. 3. The time series plots for the reactant A feed flow (variable 1), the reactant A feed valve (variable 44), the reactor level (variable 8), and the reactor pressure (variable 7). Fault 6 occurs at t = 24 hr.

To evaluate the outlier detection algorithms, 960 normal data and 480 Fault 6 data were generated. The outlier detection algorithms were used to identify the most consistent 720 observations (half of the total samples). The performance was evaluated in terms of the number of correctly identified normal data in those 720 observations.

4. RESULTS AND DISCUSSION

4.1. Outlier detection

For the original RHM algorithm, it is suggested that the upper 5% (cutoff point = c = 0.95) of the vector lengths should be checked for outliers. It is important to note that the cutoff point is correlated with the number of outliers in the data set. For a data set with large numbers of outliers, a lower cutoff point is desired. For c = 0.95, only the most extreme outliers

were identified. As c decreased to 0.75, RHM detects more outliers. As c decreases to 0.5, the swamping effect is observed. A tuning procedure is required in order to determine the optimal cutoff point for a given data set.

One way to determine the optimal cutoff point is to plot the histogram of the vector lengths from all resampling experiments (see Fig. 4). A cutoff point can be chosen as the point in which two distinct distributions are seen. For Fault 6 data, a cutoff point corresponding to a vector length of 8 would appear optimal and 93.3% were correctly identified as normal data.

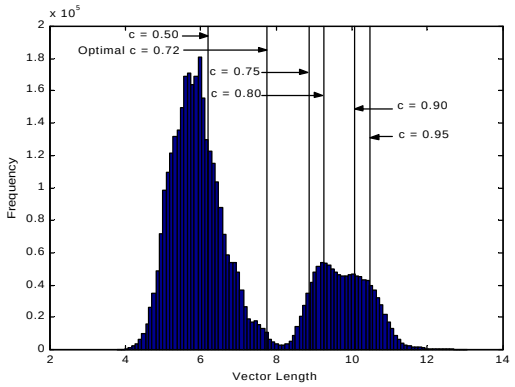

Fig. 4. The histogram of the vector lengths for the normal and Fault 6 data using RHM.

All versions of SHV correctly identify the normal data for more than 99% of the observations. The motivation to use CDC_2 or CDC_m is that they are conceptually similar to SHV, and the computation time is far less. For a data set with n observations, it is required to compute $n(n\text{-}1)/2$ Euclidean distances for SHV, versus n Euclidean distances for CDC. In other words, CDC runs $(n\text{-}1)/2$ times faster than SHV. The saving in computation time is significant when n is large.

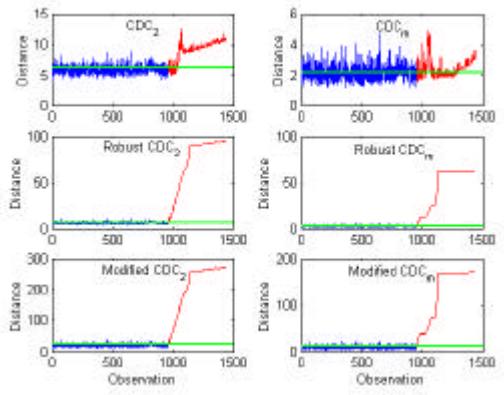

Fig. 5. The distances using CDC. Observations 1-960 represent normal data and observations 961-1440 represent Fault 6 data. Solid line represents median of the distance.

Fig. 5 displays the Euclidean distances and maximum-norm distances for CDC_2 and CDC_m, respectively. Robust CDC_2, robust CDC_m, modified CDC_2, and modified CDC_m all resulted in 100% success rate in identifying normal data. This suggests that all outliers are far away from the median and that it is a good measure to identify normal data based on the nearest distances to the median for Fault 6 data. While CDC_2 and CDC_m are able to identify the majority of the normal data, they are far less sensitive than the robust and modified version of CDC. This indicates that the mean of all of the observations is different than the mean of the normal data and the outliers have disrupted the estimation of the true mean of the normal data.

The initial step of MVT requires computation of the Mahalanobis distance, which is found to be an ineffective step for identifying outliers. This is shown in Fig. 6, in which the Mahalanobis distances are plotted for MVT, robust MVT, and modified MVT after the first and tenth iterations. For the first iteration of MVT, the half of the total observations with the smallest Mahalanobis distances were contaminated with outliers, further iterations did not improve the proficiency of MVT. For robust MVT and modified MVT, the half of the total observations with the smallest Mahalanobis distances contains mainly normal data. In this situation, further iterations do improve the proficiency of MVT.

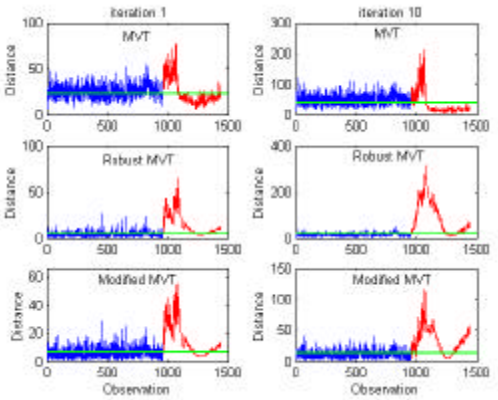

Fig. 6. The distances using MVT for iterations 1 and 10. Observations 1-960 represent normal data and observations 961-1440 represent Fault 6 data. Solid line represents median of the distance.

Robust CDC_m and modified CDC_m result in an accurate estimation of the mean and covariance of the normal data. Further iterations improve the proficiency of MVT slightly.

4.2 Fault detection and identification

Fault 6 occurs at t =24 hr. For time period 24-29 hr, GA/FDA selects the reactant A feed flow (variable 1) 99 times (see Fig. 7), indicating that this variable is strongly related to the root cause of Fault 6. The optimal fitness function $f_{GA/FDA,opt}$ is 0.993 (corresponds to 100% correct in cross validated classification result) when a single variable, reactant

A feed flow, is selected. At the same time period, the T^2 statistic contribution chart indicates that the reactant A feed valve (variable 44) contributes the most to Fault 6 and the Q statistic contribution chart indicate that the reactant A feed flow (variable 1) contributes the most to Fault 6 (see Fig. 7). Using GA/FDA provides more direct indication for the root cause.

For time period 29-34 hr, Fault 6 propagates to more than half of the total variables in the process. It will be more difficult to identify the root cause as the number of affected variables increases. As shown in Fig. 7, the reactant A feed flow (variable 1) is still selected the most by GA/FDA, although the frequency of selection decreases to 18. The contribution charts indicate that the stripper pressure (variable 16), the reactor cooling water valve (variable 51), the stripper steam valve (variable 19), and the separator pressure (variable 13) contribute the most to Fault 6 at this time period (see Fig. 7). Time series plots for these four variables show that significant step changes are found. While contribution charts detect changes in the variables for Fault 6, this does not directly lead to diagnosing the root cause (Loss of component A in feed stream 1)

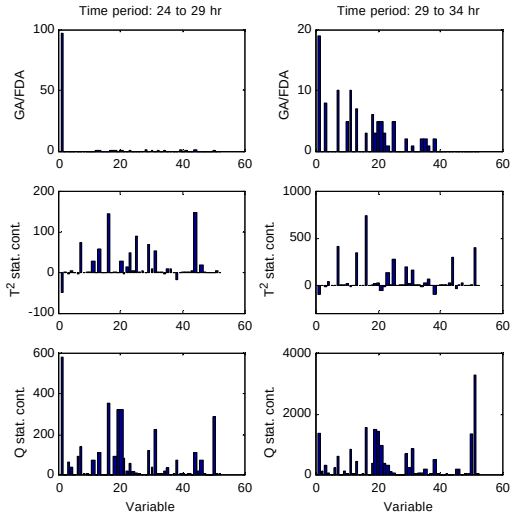

Fig. 7. The variable selection using GA/FDA, the T^2 statistic contribution chart, and the Q statistic contribution chart for the period between 24-29 hr (left hand side of the plot) and between 29-34 hr (right hand side of the plot).

5. CONCLUSIONS

To extract normal data from a historical database, robust outlier detection algorithms such as RHM, SHV, MVT, and CDC can be used. Using CDC as an initial estimate in MVT results in the best overall results using the Tennessee Eastman process data. Modified scaling is more sensitive in detecting outliers.

GA/FDA correctly identifies the variables that are responsible for the root causes for the TEP data. For cases where the process fault propagates downstream and affects more variables, GA/FDA has a better persistence in identifying the root causes as compared to contribution chart

REFERENCES

Beebe, K. R., R. J. Pell and M. B. Seasholtz (1998). *Chemometrics: A Practical Guide*, John Wiley & Sons.

Chiang, L. H., E. L. Russell and R. D. Braatz (2001). *Fault Detection and Diagnosis in Industrial Systems*, Springer-Verlag.

Chiang, L. H., R. J. Pell and M. B. Seasholtz (2003). Exploring process data with the use of robust outlier detection Algorithms. *J. of Process Control*, (in press)

Downs, J. J and E. F. Vogel (1993). A plant-wide industrial process control problem. *Comp. & Chem. Engr.*, **17,** 245-255

Duda, R. O. and P. E. Hart (1973). *Pattern Classification and Scene Analysis*, John Wiley & Sons

Egan, W. J. and S. L. Morgan (1998). Outier detection in multivariate analystical chemical data. *Anal. Chem.*, **70,** 2372-2379.

Georgakis, C., B. Steadman and V. Liotta (1996). Decentralized PCA Charts for performance assessment of plant-wide control structures. In *Proc. of the 13th IFAC World Congress***,** 97-101, IEEE Press, New Jersey

Huber, P. (1989). *Robust Statistics*, John Wiley & Sons

Jackson, J. E. (1959). Quality control methods for several related variables. *Technometrics*, **1,** 359-377.

Jackson, J. E. and G. S. Mudhlkar (1979). Control procedure for residuals associated with principal component analysis. *Technometrics*, 2**1,** 341-349

Leardi, R. (2001). Genetic algorithms in chemometricsand chmistry: a review. *J. of Chemometrics*, **15,** 559-569.

Leardi, R., R. Boggia and M. Terrile (1992). Genetic algorithms as a strategy for feature selection. *J. of Chemometrics*, **6,** 267-281.

Lyman, P. R. and C. Georgakis (1995). Plant-wide control of the Tennessee Eastman problem. *Comp. & Chem. Engr.*, **19,** 321-331.

MacGregor, J. F and T. Kourti (1995). Statistical process control of multivariate process. *Control Engr. Practice*, **3,** 403-414.

Miller, P. and R. E. Swanson (1998). Contribution plots: a missing link in multivariate quality control. *Appl. Math. and Comp. Sci..*, **8,** 775-792.

Pearson, P. K. (2001). Exploring process data. *J. of Process Control*, **11,** 179-194.

Wachs, A and D. R. Lewin. (1999). Improved PCA methods for process disturbance and failure identification. *AIChE J*, **45,** 1688-1700.

Walczak, B. and D. L. Massart (1995). Robust principal components regression as a detection tool for outliers. *Chemom. Intell. Lab. Syst.*, **27,** 41-52