

## FAULT DETECTION AND DIAGNOSIS USING CORRELATION COEFFICIENTS BETWEEN VARIABLES

MAK WENG YEE<sup>1</sup> & KAMARUL ASRI IBRAHIM<sup>2\*</sup>

**Abstract.** Chemical plants have become increasingly complex and stringent requirements are needed on the desired final product quality. Accurate process fault detection and diagnosis (PFDD) at an early stage of the process is important to modern chemical plants to achieve the above requirements. This paper focuses on the application of fault detection and diagnosis using correlation coefficients between process variables as a PFDD tool. An industrial distillation column is modelled and chosen as the case study. Principal Component Analysis (PCA) and Partial Correlation Analysis (PCorrA) are used to develop the correlation coefficients between the process variables and selected quality variables of interest. Faults considered in this research are sensor faults, valve faults and controller faults. These faults are comprised of single cause faults and multiple cause faults as well as significant faults and insignificant faults. Shewhart Control Chart and Range Control Chart are used with the developed correlation coefficients to detect and diagnose the pre-designed faults in the process. Results show that both methods based on PCA and PCorrA have good PFDD performance. In this study, the PCorrA method was better than the PCA method in detecting insignificant faults.

**Keywords:** Multivariate statistical process control; principal component analysis; partial correlation analysis; fault detection and diagnosis; correlation coefficients

**Abstrak.** Operasi loji kimia pada masa kini menjadi semakin kompleks dan kawalan kualiti yang ketat pada produk akhir diperlukan. Pengesanan dan diagnosis sebarang kecacatan dalam proses dengan cepat adalah penting bagi sesebuah loji kimia untuk mencapai kualiti produk yang diinginkan. Kertas kerja ini memfokus kepada aplikasi pengesanan dan mendiagnosis menggunakan pekali-pekali korelasi antara pemboleh ubah proses sebagai alat pengesanan dan mendiagnosis kecacatan proses. Sebuah kolom penyulingan dari industri dimodelkan sebagai kes kajian penyelidikan ini. Analisis Komponen Prinsipal (PCA) dan Analisis Korelasi Separa (PCorrA) digunakan untuk menerbitkan pekali korelasi antara pemboleh ubah proses dengan pemboleh ubah kualiti pilihan yang dikaji. Kecacatan proses yang dikaji merangkumi kecacatan injap, pengesanan dan pengawal. Kecacatan-kecacatan ini terdiri daripada kecacatan punca tunggal, kecacatan punca pelbagai, kecacatan besar dan kecacatan kecil. Carta Kawalan Shewhart dan Carta Kawalan Julat digunakan bersama dengan pekali-pekali korelasi yang diterbitkan untuk pengesanan dan diagnosis kecacatan-kecacatan yang dimasukkan ke dalam proses. Keputusan menunjukkan kedua-dua kaedah berasaskan PCA dan PCorrA boleh mengesan dan mendiagnosis kecacatan dalam proses. Dalam penyelidikan ini, kaedah PCorrA adalah lebih baik berbanding kaedah PCA dalam pengesanan dan diagnosis kecacatan-kecacatan kecil.

**Kata kunci:** Proses kawalan multipemboleh ubah statistik; analisis komponen prinsipal; analisis korelasi separa; pengesanan dan diagnosis kecacatan; pekali-pekali korelasi

---

<sup>1&2</sup>Department of Chemical Engineering, Faculty of Chemical and Natural Resources Engineering, 81310 UTM Skudai, Johor Bahru, Malaysia

\* Corresponding author: Tel: 0197167000. Email: kamarul@fkkksa.utm.my

## 1.0 INTRODUCTION

Multivariate Statistical Process Control (MSPC) is a suitable method for application in chemical industries which are multivariable in nature. MSPC monitoring method consists of collecting nominal operation condition (NOC) process data, building process models using multivariate projection methods such as Principal Component Analysis (PCA) and Partial Least Squares Analysis (PLS) and comparing the incoming online process measurements against the developed process models. According to Yoon and MacGregor [1], MSPC is a powerful tool for fault detection but its main limitation lies in the ability to diagnose the actual causes of detected faults. Conventional contribution plots used to diagnose the faults tend to be noisy and ambiguous because these plots do not have confidence limit. The determination of whether a situation is normal or abnormal lies on the judgment of the plant operators.

Lee *et al.* [2] try to improve the process fault detection and diagnosis ability of MSPC by using Independent Component Analysis (ICA) in the place of PCA. ICA focuses on extracting latent variables (independent variables in process that are not directly measurable) from the process data. These latent variables are then used with conventional MSPC process monitoring techniques. The results show that the developed ICA based method was able to detect the pre-designed faults. The process fault diagnosis using contribution plots is ambiguous since there is no limit to differentiate the situations of fault or normal process operation. Choi and Lee [3] proposed using contribution plots with limits for fault diagnosis to overcome the previous ambiguous nature of contribution plots. The introduction of relative contribution in contribution plots with limits provides a more effective and confident fault diagnosis of detected faults. The limitation of this method is the limit of the contribution plots is selected arbitrarily and not by any statistical or empirical method, thus affecting the validity of the limit selected. The present research work focuses on developing a process fault detection and diagnosis algorithm that can overcome the ambiguous nature of fault diagnosis in conventional MSPC monitoring techniques.

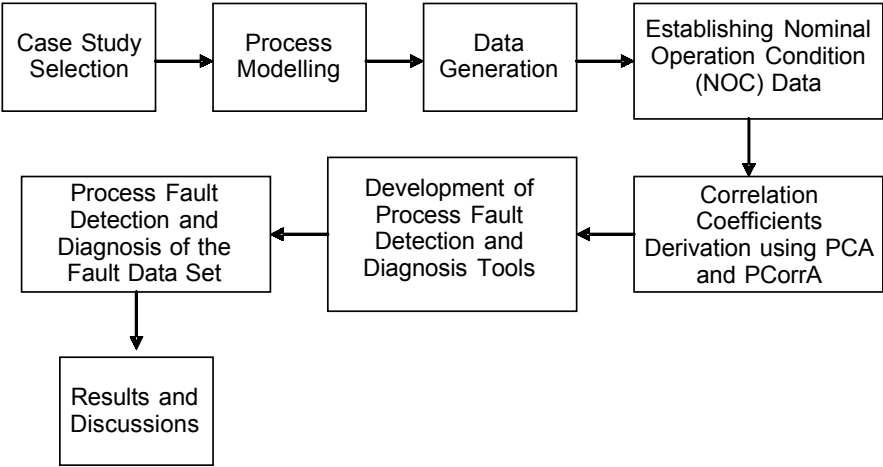
## 2.0 METHODOLOGY

The methodology of this research work is summarized as shown in Figure 1.

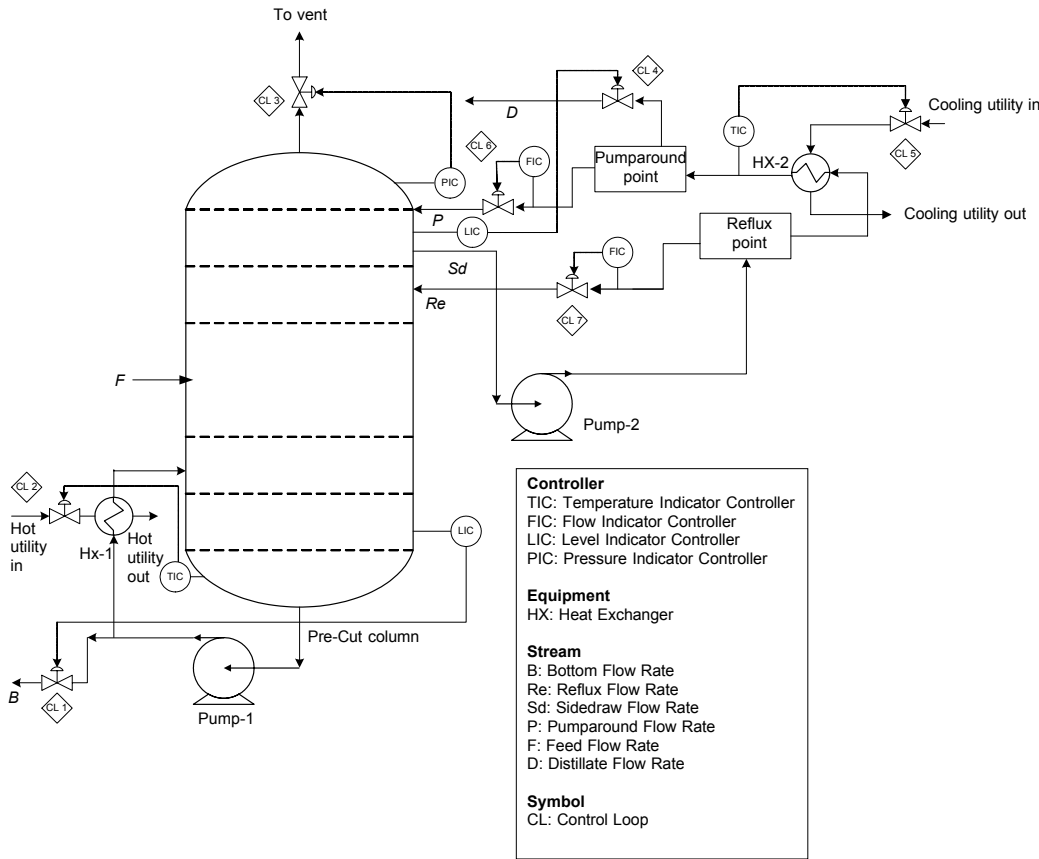
### 2.1 PROCESS MODELLING AND DATA GENERATION

A distillation column from a Palm Oil Fractionation Plant developed by Wong [4] with slight modifications is selected as the case study of this research. The product of the column are the bottom stream with oleic acid and linoleic acid composition in this product stream chosen as the quality variables of interest in this research. Figure 2 shows the distillation column with process variables and control loops of the process.

From Figure 2, there is one input stream, the feed stream and three output streams: distillate stream, bottom product stream and a vent stream. Controllers are installed



**Figure 1** Flowchart of methodology of research work



**Figure 2** Distillation column model

in the column for steady operation as shown in Figure 2. The state equations for the distillation column were derived from first principal equations. Ordinary Differential Equations (ODE) for state equations were formed and solved using 4<sup>th</sup> Order Runge-Kutta method with a step size of 0.005. The MATLAB software was used for the whole simulation program. From the column model, two sets of process operating data were generated: Nominal Operation Condition (NOC) data and Out of Control (OC) data. NOC refers to conditions when the selected key process variables and quality variables of interest remain close to their desired values. The selected key process variables are feed flow rate, feed temperature, reboiler duty, reflux flow rate, pumparound flow rate and bottom column temperature. These variables are chosen based on their normal correlation value with the two quality variables of interest, oleic acid and linoleic acid composition (all these variables shown an absolute normal correlation value of more than 0.1 with the quality variables of interest with the value of  $\geq 0.1$  selected as threshold value for selecting key process variables). For NOC data, some noises with zero mean were imbedded into the simulation program; these noises are small random change in process variables such as feed flow rate, feed temperature, reboiler duty, cooler duty, reflux flow rate and pumparound flow rate. For OC data, some large changes (significant faults) and moderate changes (insignificant faults) were added into the process model as faults. In NOC data, the two quality variables have a variation of  $\pm 3\sigma$  (in statistics, this common cause situation is where out of 1000 data points, there are only 3 data points that have values out of range [5]), significant faults represents situations where these two quality variables have values of exceeding  $\pm 4\sigma$  while insignificant faults represents situations between  $\pm 3\sigma$  and  $\pm 4\sigma$ . Process faults represent valve faults, sensor faults and controller faults. The description of each type of fault is described in Table 1 [6].

## 2.2 ESTABLISHING NOMINAL OPERATION CONDITION (NOC) DATA

In the NOC data set, the values of the selected key process variables have variations such that the two quality variables of interest, oleic acid and linoleic compositions in the bottom stream,  $y_1$  and  $y_2$ , have variation of  $\pm 3\sigma$ . 50 data points with the stated conditions for  $y_1$  and  $y_2$  and selected key process variables are used in the NOC data set. The NOC data are mean-centered and variance-scaled, then subjected to analysis using PCA and PCorrA for deriving the correlation coefficients between the selected key process variables and  $y_1$  and  $y_2$ .

## 2.3 DERIVATION OF CORRELATION COEFFICIENTS

Correlation coefficients represent the relationship between the selected key process variables, feed flow rate ( $L_f$ ), feed temperature ( $T_f$ ), reflux flow rate ( $Re$ ), pumparound flow rate ( $P$ ), reboiler duty ( $Q_r$ ), bottom column temperature ( $T_{bot}$ ) and  $y_1$  and  $y_2$ . Correlation coefficients are derived using PCA and PCorrA.

**Table 1** Fault descriptions

| (1) Sensor fault  | (2) Valve fault   | (3) Controller fault   |
|---|---|--|
| <ul style="list-style-type: none"> <li>For open loop variables, only the value of the variable changes abnormally. For closed loop variables, only the value of the disturbance( D ) OR the manipulated variable (MV) OR the control variable (CV) changes abnormally.</li> </ul> | <ul style="list-style-type: none"> <li>For open loop variables, only the value of the variable changes abnormally. For closed loop variables, both the value of manipulated variable (MV) AND control variable (CV) changes abnormally together.</li> </ul> | <ul style="list-style-type: none"> <li>For closed loop variables, the value of manipulated variable (MV) AND control variable (CV) changes abnormally together.</li> </ul> |
|   |   |  |

### 2.3.1 Correlation Coefficients Using PCA

Method for obtaining correlation coefficients between the variables,  $C_{ik}$ , using PCA was based on the research work by Lam and Kamarul [7]. Correlation coefficients using PCA are calculated as in Equation 1.

$$C_{ik} = \sum_{j=1}^n v_{ij} v_{kj} \lambda_j \quad (1)$$

Where:  $C_{ik}$  = correlation coefficient between key process variable  $i$  and quality variable  $k$

$v_{ij}, v_{kj}$  = eigenvectors obtained from NOC data using PCA  
 $\lambda_j$  = eigenvalue obtained from NOC data using PCA

### 2.3.2 Correlation Coefficients Using PCorA

PCorA determines the correlation between two variables while allowing the effect of other correlated variables on these two variables. For calculating correlation coefficient,  $C_{ik}$ , for  $y_1$  and  $L_f$  using PCorA after allowing the effect of  $j - 2$  ( $j$  represent the total number of variables, in this research,  $j = 7$  as there are 6 variables to be determine their  $C_{ik}$  values with each quality variables of interest (two quality variables of interest)) variables is as shown in Equation 2 [8].

$$C_{y_1 L_f} = \frac{r_{12(4, \dots, j-2)} - r_{13(4, \dots, j-2)} r_{23(4, \dots, j-2)}}{\left(1 - r_{13(4, \dots, j-2)}^2\right)^{1/2} \left(1 - r_{23(4, \dots, j-2)}^2\right)^{1/2}} \quad (2)$$

Where:  $C_{y_1 L_f}$  = correlation coefficient between  $y_1$  and  $L_f$  after the effect of  $j - 2$  variables

$r_{12.3}$  = partial correlation between key process variables 1 and 2 after the effect of key process variable 3

### 2.3.3 Differences between PCA and PCorA in Correlation Coefficient Derivation

PCA determines the correlation between two variables while taking account into the cross-correlation between these two variables with other variables. The derived correlation coefficient represents the correlation between the two variables with incorporation of variation information of the two variables with other correlated variables. The effects of variation of other variables on these two variables are not omitted during the derivation of correlation. In PCorA, the effects of variation of other variables with the two variables are controlled when determining the correlation between the two variables. The correlation derived is 'pure', free of effects of variation of other correlated variables with the two focused variables.

## 2.4 DEVELOPMENT OF PROCESS FAULT DETECTION AND DIAGNOSIS TOOLS

The selected key process variables are represented as  $x_i$  (with  $i = 1, \dots, 6$ ) and the quality variables of interest are represented as  $y_i$  (with  $i = 1, 2$ ).  $C_{ik}$  relates a key process variable,  $x_i$ , with a quality variable of interest,  $y_i$ , in the following way:

$$x_i = \frac{y_i}{C_{ik}} \quad (3)$$

For conventional Shewhart Control Chart, the Upper Control Limit (UCL), Center Line (CL) and Lower Control Limit (LCL) for mean-centered and variance-scaled variables are +3, 0 and -3 respectively [5]. These values are used because the NOC data were determined using a  $\pm 3\sigma$  variation in the two quality variables as threshold value for NOC condition. Using the information from Equation 3, the UCL, CL and LCL for quality variables and selected key process variables will be +3, 0 and -3 and  $+3/C_{ik}$ , 0 and  $-3/C_{ik}$  respectively. The UCL, CL and LCL for conventional Range Control Chart for mean-centered and variance-scaled variables are mean of the range values,  $R_{mean}$  multiplied by a constant,  $d_2$ ,  $R_{mean}$  and 0 respectively [5]. The constant,  $d_2$ , is determined by the number of subgroup used in calculating the range values. In the present work,  $d_2$  is 3.267 for a subgroup of  $n = 2$  [9].  $R_{mean}$  is determined by using Equation 4.

$$R_{mean} = \frac{\sum_{i=1}^n R_i}{n} \quad (4)$$

Where:  $R_i$  =  $i$ -th range value  
 $R_{mean}$  = mean of the range values  
 $n$  = number of mean values

The UCL, CL and LCL for the Range Control Chart of quality variables will be of the conventional Range Control Chart. For the selected key process variables, the UCL, CL and LCL will be  $(R_{mean} * d_2)/C_{ik}$ ,  $(R_{mean})/C_{ik}$  and 0 respectively. The developed Shewhart Control Charts and Range Control Charts will be used as process fault detection and diagnosis tools for the OC data set.

## 2.5 PROCESS FAULT DETECTION AND DIAGNOSIS OF OC DATA

The major assumption in the proposed method is that all selected key process variables and quality variables of interest are measured. The selected key process variables that are major contributors to the variation of the quality variables of interest are included into the correlation analysis. In this manner, the behavior of the process (quality variables of interest) will be well represented by the correlation determined from the NOC data and the developed PFDD algorithm method will suit the dynamic behavior of the process. The feed compositions, feed flow rate and feed temperature to the column is assumed fixed at certain values, any change in the value of feed flow rate and feed temperature are considered due to sensor fault or valve fault in

the feed stream. From Figure 2, the study column is installed with several control loops to ensure the stable operation of the column. Any common cause changes in the column either through load problem (disturbance changes) or servo problem (set point changes) will be taken care of through the installed controllers. The causal changes of interest in this work are those involving abnormal changes in the values of the selected key process variables and quality variables of interest through faults in sensors, valves or even controllers, not the two mentioned problems in the previous sentence.

When a key process variable or quality variable deviate from its NOC value, the variable of the control chart will be checked whether is a closed loop variable (closed loop variables are variables involved in controller loops) or open loop variable (open loop variables are variables not involved in controller loops). A fault signal is observed only when either the Range Control Chart (RCC) or Shewhart Control Chart (SCC) of one or both quality variables of interest show value that exceeds its control limit AND one or more selected key process variable observed a value out of its control limit either in its RCC or SCC. For open loop variables, the fault type will be of sensor or valve fault as pre-designed while faults for closed loop variable can be of valve, sensor or controller faults.

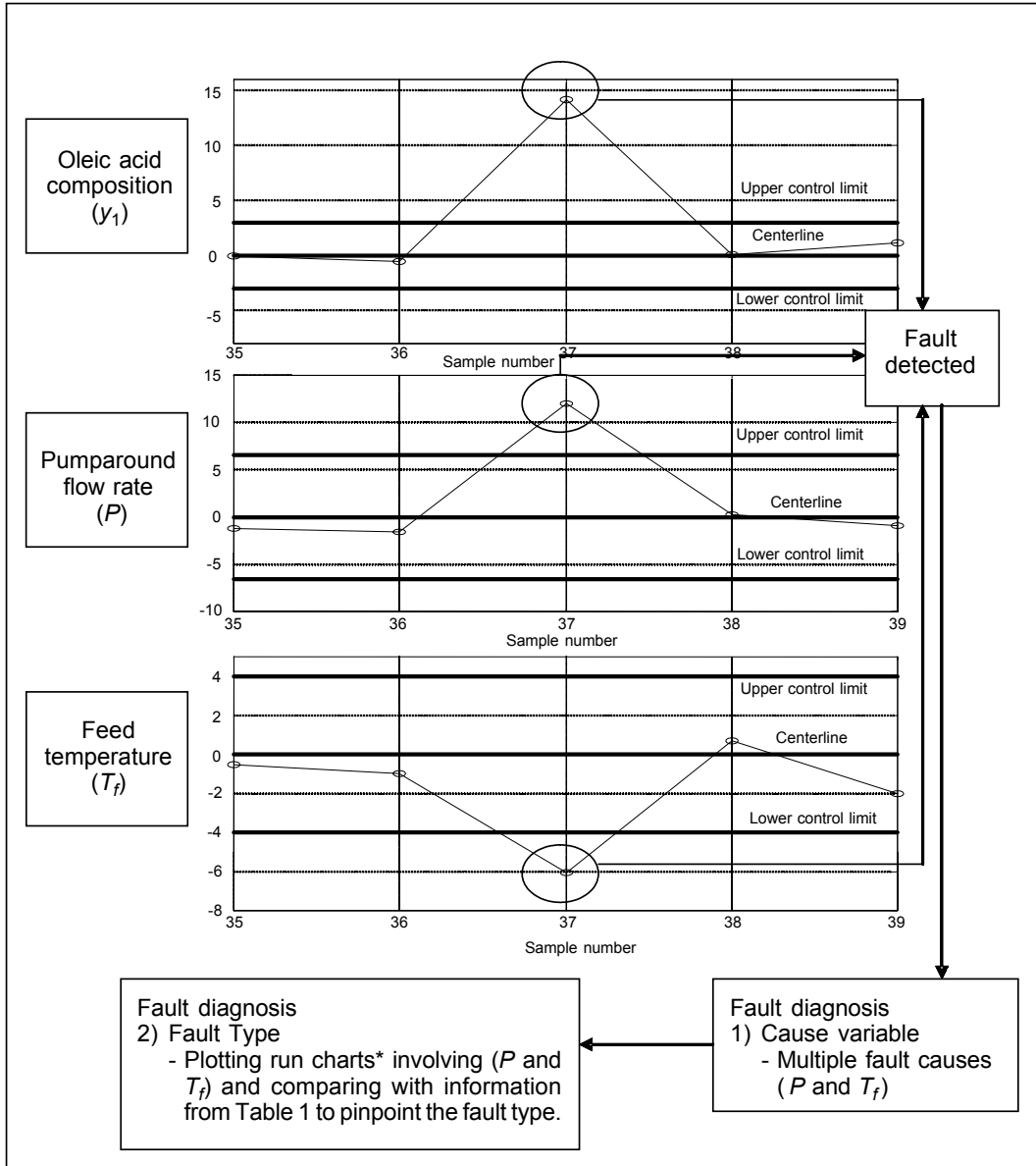
The cause variable (s) of each detected fault is diagnosed by checking the control charts of the selected key process variables. Selected key process variables that show value exceeding its control limits (either in SCC or RCC) are diagnosed as the cause of the detected fault. To determine exactly the type of detected fault, the method described in the previous paragraph is applied.

### 3.0 RESULTS AND DISCUSSIONS

Figure 3 shows an example of fault detection and diagnosis (FDD) using the developed method based on PCA for Shewhart Control Chart for a multiple causes, significant, fault. For the PCorrA method, a similar method of detection and diagnosis is used. The procedure for Range Control Chart in detecting and diagnosing faults is similar to that of Shewhart Control Chart. There were 17 pre-designed faults in the OC data set: 13 significant faults (single and multiple cause faults) and 4 insignificant faults (single and multiple cause faults). The FDD algorithm based on PCA successfully detected and diagnosed 13 faults with the 4 insignificant faults not detected. The FDD algorithm based on PCorrA successfully detects and diagnosed all the 17 pre-designed faults.

Figure 4 shows how the method based on PCorrA managed to detect and diagnose an insignificant fault while the method based on PCA failed to do so. Figures 5 and 6 shows the performance of the methods based on PCA and PCorrA in detecting faults and diagnosing the fault causes.

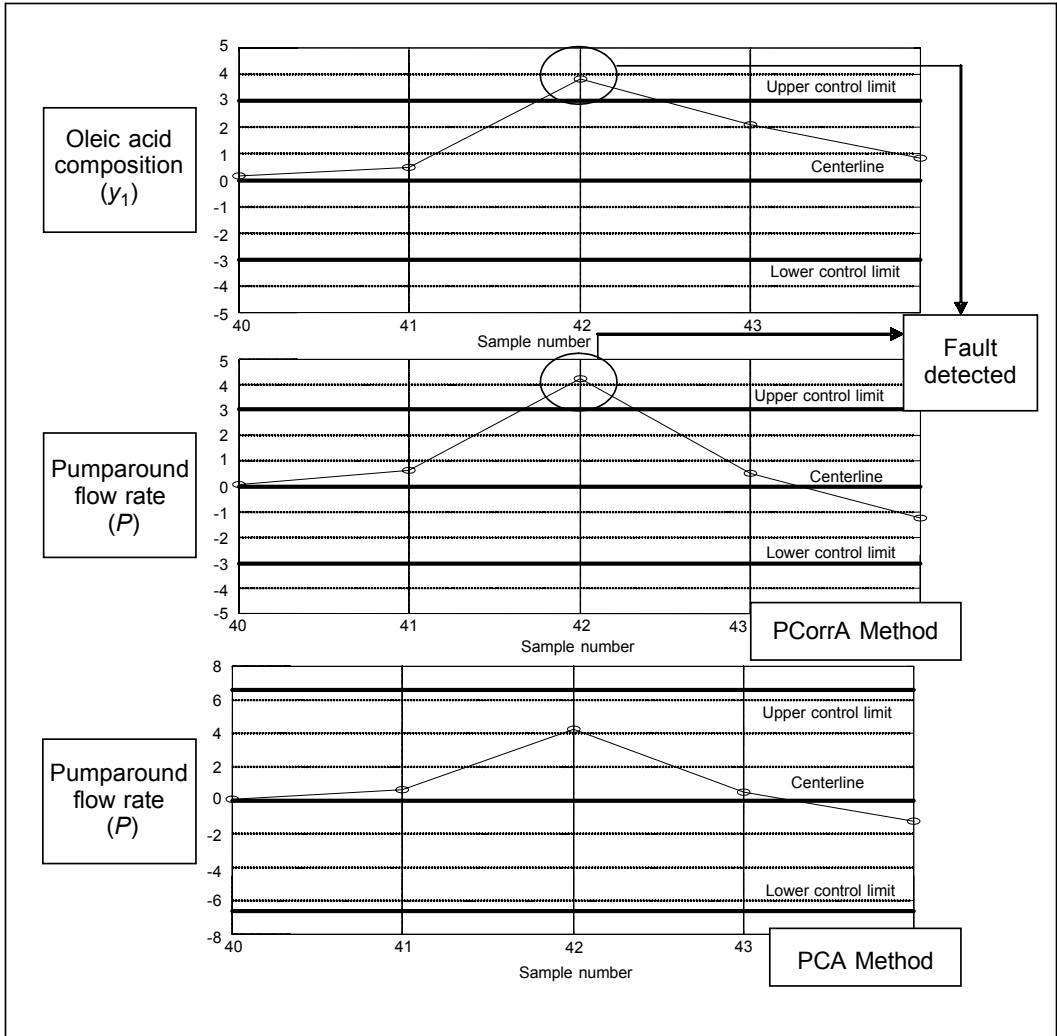
The PCorrA method performed better because the correlation coefficients developed by this method are closer to the actual value of the correlation coefficients



\*Run charts are charts that plot the values of a variable over time

**Figure 3** Example of fault detection and diagnosis based on PCA method

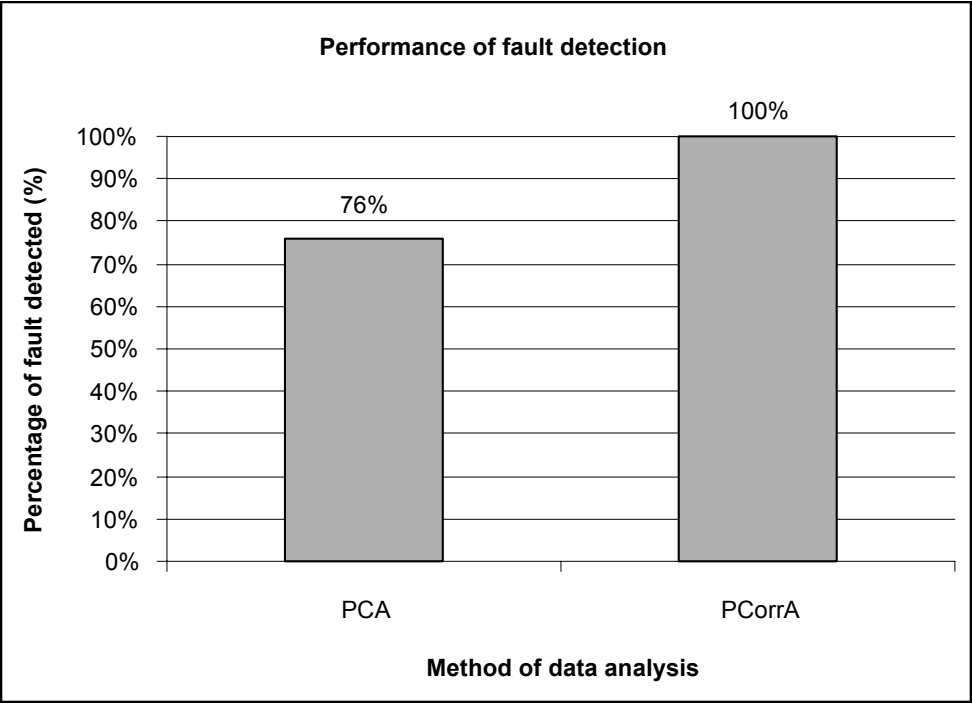
representing the correlation between the selected key process variables with the two quality variables of interest. This is because the PCorA method sets other selected key process variables at constant values when calculating the correlation between a selected key process variable with a quality variable. This feature of omitting the variation between the other key process variables with the two variables (one key



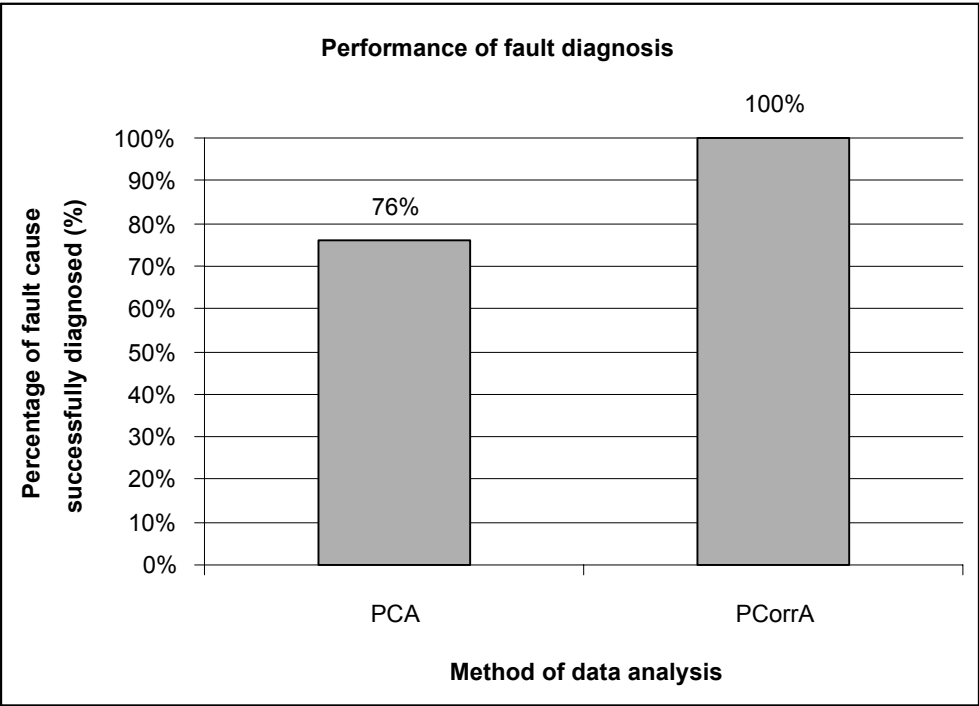
**Figure 4** Example of PCCorA method in detecting an insignificant fault while PCA method could not detect the insignificant fault

process variable and one quality variable) is a significant advantage of PCCorA over PCA in analyzing highly correlated multivariate data.

One major advantage of the developed FDD algorithm using correlation coefficients is the simplicity in determining the fault cause(s) of a detected fault. The control charts of the selected key process variables will trigger alarm if any of them exhibit value out of their control limits and charts that trigger an alarm will be determined as the root causes of the detected fault. Furthermore, the availability of control limits in these control charts will shed away any ambiguities of whether a change in value of the selected key process variables are due to common cause (NOC) or causal



**Figure 5** Performance of fault detection



**Figure 6** Performance of fault diagnosis

cause (OC). For on-line process monitoring, the data set used for calculating the correlation coefficients can be updated with dynamic data to take account into changes of the process due to change in raw material, fouling in heat exchangers and other changes in the process parameters. This area can be further researched and serve as a research problem for future work.

#### 4.0 CONCLUSION

Fault detection and diagnosis (FDD) using correlation coefficients based on PCA and PCorrA were presented. The performance of the approach was studied on an industrial distillation column. The results show that the FDD method using correlation coefficients was able to detect the pre-designed faults and diagnose the fault cause of each detected fault (both single cause faults and multiple cause faults). PCorrA managed to detect all the faults (both significant and insignificant faults) while PCA only managed to detect the significant faults. This is due to the fact that PCorrA determines the correlation between two variables after omitting the effects of other variables that are correlated with the two variables. Therefore, the correlation coefficients developed using PCorrA method was better in representing the correlation between the selected key process variables and the quality variables of interest.

#### 5.0 NOTATION

|                  |   |
|------------------|---|
| $C_{ik}$         | : Correlation coefficient                       |
| CL               | : Centerline                                    |
| CV               | : Control variable                              |
| D                | : Disturbance                                   |
| LCL              | : Lower Control Limit                           |
| $L_f$            | : Feed flow rate                                |
| MSPC             | : Multivariate Statistical Process Control      |
| MV               | : Manipulated variable                          |
| NOC              | : Nominal Operation Condition                   |
| OC               | : Out of Control                                |
| ODE              | : Ordinary Differential Equation                |
| PCA              | : Principal Component Analysis                  |
| PCorrA           | : Partial Correlation Analysis                  |
| $P$              | : Pumparound flow rate                          |
| $Re$             | : Reflux flow rate                              |
| $T_f$            | : Feed temperature                              |
| UCL              | : Upper Control Limit                           |
| $v_{ij}, v_{kj}$ | : eigenvectors obtained from NOC data using PCA |
| $x$              | : Selected key process variable                 |

- $y$  : Quality variable of interest  
 $\lambda_j$  : eigenvalue obtained from NOC data using PCA

## ACKNOWLEDGEMENTS

The present work was funded by the National Science Fellowship (NSF) scholarship. Special thanks to Mr. Wong Teck Siang for providing the base model of the case study of this work. Mr. Lam Han Loong's help in providing the information on the derivation of the correlation coefficient via eigenvector-eigenvalue approach for the PCA method is gratefully acknowledged.

## REFERENCES

- [1] Yoon, S. K. and J. F. MacGregor. 2001. Fault Diagnosis with Multivariate Statistical Models Part I: Using Steady-state Fault Signatures. *Journal of Process Control*. 11: 387-400.
- [2] Lee, J. M., C. K. Yoo and J. B. Lee. 2004. Statistical Process Monitoring with Independent Component Analysis. *Journal of Process Control*. 14: 467-485.
- [3] Choi, S. W. and I. B. Lee. 2005. Multiblock PLS-based Localized Process Diagnosis. *Journal of Process Control*. 15: 295-306.
- [4] Wong, T. S. 2003. *Advanced Process Control of a Fatty Acid Distillation Column Using Artificial Neural Network*. Masters Thesis (Universiti Teknologi Malaysia), Malaysia.
- [5] McNeese, W. H. and R. A. Klein. 1991. *Statistical Methods for the Process Industries*. USA: ASQC Quality Press.
- [6] Mak, W. Y. and K. A. Ibrahim. 2003. Fault Detection of A Distillation Column Using Multivariate Statistical Process Control. *Proceedings of The 17<sup>th</sup> Symposium of Malaysian Chemical Engineers (SOMChE 2003), 29 – 30 December 2003, Copthorne Hotel, Penang, Malaysia*. 92 – 96.
- [7] Lam, H. L. and K. A. Ibrahim. 2002. Improved Multivariate Statistical Process Control for Chemical Process Fault Detection and Diagnosis via Cross Variable Correlation Approach. *Regional Symposium on Chemical Engineering, RSCE/SOMChem 2002, Hotel Hilton, Petaling Jaya, 28 – 30 October 2002*. 2: 1637-1644.
- [8] Cliff, A. D. and J. K. Ord. 1973. *Spatial Autocorrelations*. London: Pion.
- [9] Wetherill, G. B. and D. B. Brown. 1991. *Statistical Process Control: Theory and Practice*. Chapman and Hall.