Available online at www.sciencedirect.com
SCIENCE DIRECT®
Control Engineering Practice 12 (2004) 745–755
CONTROL ENGINEERING PRACTICE
www.elsevier.com/locate/conengprac

Synthesis of T^2 and Q statistics for process monitoring

Q. Chen[a,*], U. Kruger[b], M. Meronk[c], A.Y.T. Leung[d]

[a] College of Aerospace Engineering, Nanjing University of Aeronautics and Astronautics, 29 Yudao Street, Nanjng 210016, China
[b] Intelligent Systems and Control Research Group, Queen's University of Belfast, BT9 5AH, UK
[c] Ruttgers Chemicals AG., Castrop Rauxel, Germany
[d] Department of Building and Construction, City University of Hong Kong, China

Received 2 July 2002; accepted 28 August 2003

Abstract

In this work, a novel approach that synthesizes the T^2 and Q statistics for statistical process condition monitoring is introduced. The synthesis is based on the joint probability density function of the two statistics, which is estimated with kernel density estimation. This inherits the following benefits: (i) it can be more sensitive to detect abnormal process behaviour than the individual statistics and (ii) it reduces the number of monitoring charts to be observed. The utility of this new approach is demonstrated using two application studies that involve the simulation of a fluid catalytic cracking unit and recorded data from an industrial distillation unit.

Keywords: Statistical process control; Data reduction; Fault detection; Process model; Probability density function; Kernel density estimation; Condition monitoring

1. Introduction

The requirement for efficiency and profit in the chemical, manufacturing and other process industries imposes a constant need for attention to condition monitoring strategies. For example, a survey by Nimmo (1995) revealed that the US-based petrochemical industry could save up to $10 billion annually if abnormal process behaviour could be detected, diagnosed and appropriately dealt with. Vedam and Venkatasubramanien (1999) highlighted that the same industry looses over $20 billion per year due to inappropriate reaction to abnormal process behaviour.

The very important task of detecting and diagnosing abnormal process behaviour has led to the evolution of a range of statistically based condition-monitoring approaches. Such approaches are collectively referred to as multivariate statistical process control (MSPC) and have gained attention over the past decade—noticeable by the large number of publications in this area (e.g. MacGregor & Kourti, 1995; Kourti & MacGregor, 1995; Morud, 1996).

MSPC aims to remove the often-observed high degree of redundancy in the data by defining a reduced set of statistically uncorrelated variables (MacGregor & Kourti, 1995; Wise & Gallagher, 1996). Principal component analysis (PCA) and partial least squares (PLS) are such approaches. They utilize orthogonal projections that exploit cross-correlation among the process variables (Wise & Gallagher, 1996; Chen, Wynne, Goulding, & Sandoz, 2000). The reduced variable set is then utilized to determine two statistics, i.e. the T^2 and Q statistics, and bivariate scatter diagrams for on-line process monitoring (MacGregor, Marlin, Kresta, & Skagerberg, 1991).

Problems that frequently arise from the application of MSPC techniques relate to (i) the assumption of normally distributed process variables and (ii) the size of the reduced variable set which may still be significant (Kruger, Chen, Sandoz, & McFarlane, 2001). More precisely, the reduced variable set may not be normally distributed although it can be shown that the variables of this reduced set are "more normally distributed". This was highlighted by Wise and Gallagher (1996) and argued on the basis of the central-limit-theorem. In addition, if the size of the reduced variable set is larger than 5 for example, the number of monitoring charts

*Corresponding author. Tel/fax: +86-25-4893221.
E-mail address: q.chen@nuaa.edu.cn (Q. Chen).

0967-0661/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.conengprac.2003.08.004

increases considerably, rendering the practical implementation of an MSPC monitoring scheme a difficult task. Using kernel density estimation (KDE) to estimate the probability density function (PDF) of the reduced variable set may overcome the problem if this set is not normally distributed (Martin & Morris, 1996; Chen, Wynne, Sandoz, & Kruger, 1998; Chen et al., 2000). However, if the reduced variable set is still large, this approach may lead to a tedious calculation phase for estimating the confidence regions of the scatter diagrams for example (Silverman, 1986).

In this article, a joint analysis of the T^2 and Q statistics is proposed to address the above deficiencies. The work, presented here, represents an in-depth analysis and an industrial exploitation of previous work by the authors (Chen, Kruger, Meronk, & Leung, 2001), where this approach was first introduced. The joint analysis gives rise to a bivariate plot in which each pair of values of the T^2 and Q statistics, corresponding to the same observation, presents a scatter point. The joint PDF is to be estimated for only two variables, which leads to a much easier estimation of the joint PDF compared to the joint PDF of a large size of reduced variable set (Simonoff, 1996) and offers only one or two simple monitoring charts for process operators. The joint PDF is estimated using KDE that allows a more accurate statistical inference if the reduced variable set is not normally distributed, and may be more sensitive as demonstrated in this article.

Further to the information of a large set of process variables being presented by a single scatter diagram, a Schewhart chart of the joint PDF is proposed to monitor process behaviour with a time base. The bivariate scatter diagram allows process trends, such as distinctive operating regimes or process drifts, to be displayed, whilst the Schewhart monitoring chart allows investigating at which time a specific event arose for instance. This is exemplified by two application studies that relate to a realistic simulation of a fluid catalytic cracking unit (FCCU) and the analysis of recorded data from an industrial distillation process.

2. Review of MSPC techniques

In this section, a brief review of PCA and PLS is presented along with the definition of their T^2 and Q statistics.

2.1. Principal conponent analysis

The PCA is a statistical modelling technique that relies on a singular value decomposition of a given data matrix $\mathbf{X}\in\mathbb{R}^{K\times L}$, which contains reference data that describe normal operating conditions of the process. K refers to the number of observations, stored as row vectors in $\mathbf{X}$, and L is the number of process variables, included in the PCA analysis. The columns of $\mathbf{X}$ are usually normalized, i.e. they are mean centred and scaled to unit variance.

The PCA decomposition of $\mathbf{X}$ is as follows:

$$\mathbf{X}=\hat{\mathbf{X}}+\mathbf{E}=\sum_{k=1}^{L}\mathbf{t}_k\mathbf{p}_k^{\mathrm{T}}+\mathbf{E}, \tag{1}$$

in which

$$\hat{\mathbf{X}}=\sum_{k=1}^{L}\mathbf{t}_k\mathbf{p}_k^{\mathrm{T}} \tag{2}$$

is the prediction of $\mathbf{X}$ on m retained principal components (PCs) and $\mathbf{E}\in\mathbb{R}^{K\times L}$ is a residual matrix, representing the PCA model prediction error. The vectors $\boldsymbol{t}_k\in\mathbb{R}^K$ are known as score vectors and describe significant process variation and $\boldsymbol{p}_k\in\mathbb{R}^L$ are the loading vectors that reveal the interrelationships between the process variables. The score vectors are given by

$$\mathbf{t}_k=\mathbf{X}\mathbf{p}_k^{\mathrm{T}}. \tag{3}$$

The division of $\mathbf{X}$ into $\hat{\mathbf{X}}\in\mathbb{R}^{K\times L}$, capturing significant variation in the data arising due to common causes, and $\mathbf{E}$, representing insignificant variation, for example measurement uncertainty and insignificant trends in the process variation, gives rise to two statistics, known as the T^2 and Q statistics:

$$T_k^2=\sum_{i=1}^{M}\frac{t_{ki}^2}{\sigma_i}, \tag{4a}$$

$$Q_k=\sum_{j=M+1}^{L}t_{kj}^2, \tag{4b}$$

where T_k^2 and $Q_k\in\mathbb{R}$ are samples of the T^2 and Q statistics at time instance k, t_{ki} and t_{kj} are the value of the ith and jth score variable, respectively, at time instance k and $\sigma_i\in\mathbb{R}$ is the estimated variance of the ith score variable. Both statistics inherit the following property that is important for the development of the new monitoring approach.

Lemma 1. *The T^2 and Q statistics are independent of each other, since the residuals of the PCA model are orthogonal to the model plane.*

Lemma 1 is proven in Appendix A. Practically, the T^2 statistic represents an overall measure of process variation, whilst the Q statistic describes how well the PCA model predicts the recorded process variables (Jackson, 1991).

To implement a PCA based process monitor on-line, the following steps are required. A PCA model, based on reference data of the process, has to be identified. This data have to be selected with care to avoid that abnormal process behaviour is retained which, in fact,

should be discarded, and to guarantee that the recorded reference set is large enough to capture normal process variation, as otherwise false alarms arise (Kruger et al., 2001). Prior to the identification of a PCA model, the reference data should be normalized. Jackson (1991) and Valle, Weihua and Qin (1996) discussed several techniques to determine the number of the retained PCs. Cross validation is found to be one of the most appropriate techniques for this purpose and is used in this work. A more detailed description of PCA may be found in (Wold, Esbensen, & Geladi, 1987; Joliffe, 1986; Jackson, 1991).

For on-line monitoring, the PCA model is applied to project a new data point onto the model plane, spanned by the retained loading vectors, and estimate the residual of the PCA model. Then, the associated values of the T^2 and Q statistics are calculated and the monitoring charts are updated. Violation of the confidence limits of either statistic is indicative of abnormal process behaviour and an alarm is raised to the process operator.

2.2. Partial least squares

PLS is designed to establish the relationships between a given predictor matrix, $\mathbf{X}\in\mathbb{R}^{K\times M}$, and a response matrix, $\mathbf{Y}\in\mathbb{R}^{K\times N}$. Both matrices are formed with the recorded data of process variables. All measured variables are divided into M cause or predictor variables that form the predictor matrix and N effect or response variables that form the response matrix. $\mathbf{X}$ and $\mathbf{Y}$ contain K observations each and their columns are typically mean centred and appropriately scaled. With PLS, the predictor and the response matrices are decomposed as follows:

$$\mathbf{X}=\hat{\mathbf{X}}+\mathbf{E}=\sum_{k=1}^{n}\mathbf{t}_k\mathbf{p}_k^{\mathrm{T}}+\mathbf{E}, \tag{5}$$

$$\mathbf{Y}=\hat{\mathbf{Y}}+\mathbf{F}=\sum_{k=1}^{n}\mathbf{t}_k\mathbf{q}_k^{\mathrm{T}}+\mathbf{F}, \tag{6}$$

where $\hat{\mathbf{X}}\in\mathbb{R}^{K\times M}$ and $\hat{\mathbf{Y}}\in\mathbb{R}^{K\times N}$ are the reconstructed predictor and response matrices respectively, $\mathbf{q}_k\in\mathbb{R}^N$ is the loading vectors of the response matrix, $\mathbf{t}_k\in\mathbb{R}^K$ and $\mathbf{p}_k\in\mathbb{R}^M$ are the score and loading vectors of the predictor matrix, $\mathbf{E}\in\mathbb{R}^{K\times M}$ and $\mathbf{F}\in\mathbb{R}^{K\times N}$ are the residual matrices of the predictor and response matrices and n the number of retained latent variables (LVs). The determination of n is usually determined by cross validation, as discussed by Kresta, MacGregor, and Marlin (1991), MacGregor and Kourti (1995) and Wise and Gallagher (1996).

The T^2 and Q statistics for PLS can be determined in a similar way to those of PCA. Eqs. (5) and (6) highlight that the PLS decomposition results in the extraction of significant process variation, described in $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, and the residual matrices $\mathbf{E}$ and $\mathbf{F}$. Since $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ are both computed using the vectors $\boldsymbol{t}_1,\ldots,\boldsymbol{t}_n$, three statistics may be established which represent (i) the variation of the t-scores variables, (ii) the residuals of the predictor variables and (iii) the residuals of the response variables at the kth time instance. These statistics are defined as follows:

$$T_k^2=\sum_{i=1}^{n}\left(\frac{t_{ki}}{\lambda_i}\right)^2, \tag{7a}$$

$$Q_k^{(Y)}=\sum_{i=1}^{N}(y_{ki}-\hat{y}_{ki})^2, \tag{7b}$$

$$Q_k^{(X)}=\sum_{j=n+1}^{M}t_{kj}^2, \tag{7c}$$

where $\lambda_i\in\mathbb{R}$ is the estimated standard deviations of the retained t-score variables. The relationship between these statistics is of importance for the development of the joint PDF and are summarized as follows.

Lemma 2. *The statistics T^2, $Q^{(Y)}$ and $Q^{(X)}$ are independent of each other.*

Lemma 2 is proven in Appendix B. The T^2 statistic represents variation within the predictor variables that considerably contributes to the prediction of the response. The $Q^{(Y)}$ statistic and the $Q^{(X)}$ statistic represents residuals of the predictor and response variables. Although successful application have been reported (Morud, 1996; Wise & Gallagher, 1996), Kruger et al. (2001) showed that the T^2 and the $Q^{(X)}$ statistics, on their own, are insufficient for process monitoring.

For process monitoring, however, only the T^2 and the $Q^{(Y)}$ statistics are considered here. This is because any changes of the correlation structure between the predictor variables implies the change in $Q^{(X)}$ statistic. If these changes are due to abnormal process behaviour then the $Q^{(Y)}$ statistic would also respond to this behaviour, which makes the $Q^{(X)}$ and $Q^{(Y)}$ statistics provide redundant information for condition monitoring. Therefore, only $Q^{(X)}$ statistic or $Q^{(Y)}$ statistic will be used for joint PDF estimation.

3. KDE preliminaries

KDE refers to a class of data driven techniques of the nonparametric estimation of PDFs (Silverman, 1986; Wand & Jones, 1995a,b). It is a powerful tool extracting an empirical distribution density from a given sample of the process population in light.

The parametric approach assumes that the form of an underline function, e.g. a PDF, is predetermined apart from its parameters. For the Gaussian density function, these parameters are the mean value and the standard deviation of the process variables. In contrast, non-parametric approaches do not need prior knowledge about the form of the density function. KDE determines the density function by summing up small "humps" that are centred to each data point. These "humps" are defined by kernel functions, e.g. Gaussian, Triangular or Epanechnikov type (Silverman, 1986).

Consider a kernel function $K(\mathbf{x})$ and a sample $\mathbf{D}_0 = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ from a population distribution density $f(\mathbf{x})$, then the density estimate of the sample can be written as

$$\hat{f}(\mathbf{x}, \mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i)), \tag{8}$$

where $\mathbf{H}$ is a smoothing or bandwidth matrix. In practice, the form of the kernel function is not of importance. In this work, the kernel functions were selected to be Gaussian kernels. Note that the accuracy of the PDF depends on the *bandwidth* or *smoothing* parameters. For determining these parameters, leave-one-out cross validation is at the heart of most bandwidth selection algorithms for KDE (Silverman, 1986).

For determining a PDF, there are several bandwidth selection algorithms to adapt to different data natures. For the well spreading data sets, mean integral squared error (MISE) cross validation is employed, which allows straightforward extensions to multivariate cases (Bowman, 1984). For a d-dimensional data set, the bandwidth parameters are given by a symmetric bandwidth matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$. To estimate all $d(d+1)/2$ elements of the $\mathbf{H}$ matrix is not a realistic task in terms of computational load and time. However, if the data is 'sphered' before being used to fit the density function i.e. re-scale the data to be the same scale in all dimensions (Fukunaga, 1990), then $\mathbf{H} = h^2\mathbf{I}_d$ is a reasonable choice for its simplicity and much reduced computational load, where h is the only bandwidth parameter and $\mathbf{I}_d$ the identity matrix of dimension d. The 'sphered' process variables are statistically independent of each other and have the same variance in all directions. Thus, there is only one parameter to be determined. Generally, the determination of $\mathbf{H}$ is based on minimising the mismatch between the "true" and estimated PDFs. Using MISE this mismatch is given by

$$MISE\{\hat{f}(:, \mathbf{H})\} = E\int [\hat{f}(\mathbf{x}, \mathbf{H}) - f(\mathbf{x})]^2 \, \mathrm{d}\mathbf{x}, \tag{9}$$

where $\hat{f}(\mathbf{x})$ is the estimated PDF and $f(\mathbf{x})$ the unknown "true" PDF and $E\{\mathbf{x}\}$ the expectation operator. Details of practical issues regarding KDE application in process monitoring may be found in Martin & Morris (1996) and Chen et al. (1998, 2000, 2001).

4. KDE on T^2 and Q statistics

In process condition monitoring and fault detection, T^2 and Q statistics are most important statistical parameters. They are depicted against time instances independently to monitor the process system performance. When the process is in normal operating condition, the variations of the T^2 and Q statistics stay within the nominal regions specified by system engineer. If anyone of the two statistics departures its nominal region systematically, the operators should be alarmed about the emerging of abnormal operating conditions. Another way to use T^2 and Q statistics is plotting them against PC/LVs to look for the PC/LV vectors affected by the faults (Kresta et al., 1991), which requires to watch all the PC/LVs detained. For *incipient* faults, the sensitivity of the two statistics is not clear. Improving the sensitivity of a monitoring chart is always desirable for process plants especially for those in which the system failure will cause catastrophic consequences or extremely costly loss. The abnormal changes in a process might be invisible in the conventional features in the early stage of a developing fault so that the early detection of the abnormal conditions is a difficult and crucial task.

As well known, the multivariate statistics has important advantages over the univariate statistics in the sense of extracting complete information of the underline data regime. Therefore, the joint statistical properties of the T^2 and Q statistics could bring out more information increasing the sensitivity to the incipient faults. Assuming the sensitivities of T^2 and Q statistics are s_1 and s_2, respectively, the sensitivity of the joint PDF of the T^2 and Q statistics could be expressed as

$$s_{\text{joint}} = as_1 + bs_2 + \delta \tag{10}$$

in which a and b are the coefficients weighing up the contribution of T^2 and Q statistics to the overall sensitivity, δ is the increased sensitivity obtained from the joint PDF. The KDE, as a new approach introduced into MSPC, has been employed to extract the empirical PDF on PCs or LVs to monitor process condition (Martin & Morris, 1996; Chen et al., 1998, 2000). This is an important development compared to the parametric method on PDF estimation.

Since the T^2 and the Q statistics are independent of each other, the joint PDF of the two statistics will be the product of two Chi-Squared density function if it can be assumed that the T^2 as well as the Q statistic have Chi-Squared distribution. Often this is not the case in practice. Particularly the T^2 may not have a Chi-Squared distribution if the process variables are not

normally distributed (Wise & Gallagher, 1996). In order to estimate the unknown density function of the joint T^2 and Q statistics, a data-driven KDE is employed and the confidence regions are obtained using an empirical method or numerical integration.

But the problem with the KDE is that if the number of PCs or LVs are more than five, the bandwidth calculation becomes a daunting task although various measures have been taken to speed up the process. Applying KDE on T^2 and Q statistics will change the scenario. KDE with only two variables will be fast, efficient and more accurate. It produces only one monitoring chart being capable of monitoring the process system performance. As both detained and discarded PC/LVs statistical behaviours have been contained in the joint PDF or its confidence regions, the small variation of the detained number of PC/LVs might not have much impact on the performance of the technique from condition monitoring point of view.

The joint PDF of the T^2 and Q statistics can be obtained using KDE:

$$\hat{f}(\mathbf{z}, h\mathbf{I}) = \frac{1}{n}\sum_{i=1}^{n} K(h^{-1/2}(\mathbf{z} - \mathbf{Z_i})), \tag{11}$$

where

$$\mathbf{z} = \begin{Bmatrix} T^2 \\ Q \end{Bmatrix}, \quad \mathbf{Z}_i = \begin{Bmatrix} T_i^2 \\ Q_i \end{Bmatrix} \text{ for PCA model,} \tag{12}$$

$$\mathbf{z} = \begin{Bmatrix} T^2 \\ SPE \end{Bmatrix}, \quad \mathbf{Z}_i = \begin{Bmatrix} T_i^2 \\ SPE_i \end{Bmatrix} \text{ for PLS model.} \tag{13}$$

In Eq. (13), SPE can be $Q^{(X)}$ statistic or $Q^{(Y)}$ statistic depending which one is used. Based on Eq. (11), a three-dimensional graph of $\hat{f}(\mathbf{z})$ can be drawn but a two-dimensional plot of confidence intervals is more useful. It can be superimposed on new on-line data points from the process. Under normal operation conditions, the new data point (T_k^2, Q_k) should land within the 99% and 95% confidence regions with occasional outliers. If the abnormal condition occurs, the data points would systematically migrate out of the 99% and 95% areas of the confident regions, or far more points than 1% and 5% of the depicted points would scatter outside the regions. This would be an indication of faulty condition of the process. The increased sensitivity of the joint PDF is illustrated in Fig. 1. In the figure, the T^2 statistic is presented against sampling time points by a univariate monitoring chart in the bottom right corner, and the Q statistic in the top left corner with empirical confidence limits. The projection of T^2 and Q statistics onto the two-dimensional plane with fitted joint PDF contours of 95% and 99% confidence regions is also shown in the top right corner of the same picture for comparison.

It is obvious that the shadowed area in the two-dimensional contour box indicates the increased sensitivity of the joint PDF over the individual statistics. Even if the process is normally distributed, the ellipse-shaped joint PDF confidence regions would still give more sensitivity over the individual statistics. Another advantage of the joint PDF approach is that the number of variables is set to 2 and hence the tedious calculation process is avoided. This also implies that the PDF and its confidence regions can be visualized without any distortion and infidelity.

It should be pointed out that although the joint PDF provides more sensitivity than the individual statistic does, the T^2 and Q statistic charts should still be utilized together with the joint PDF chart for extra information on the nature of the abnormal event corresponding to which statistic's confidence limit is violated.

5. A simulated case—FCCU

A FCCU is an important economic unit in refining operations. It typically receives several different heavy feed stocks from other refinery units and cracks these streams to produce lighter, more valuable components that are eventually blended into gasoline and other products. The particular Model IV unit described by McFarlane, Reineman, Bartee, and Georgakis (1993) is illustrated in Fig. 2.

The principal feed to the unit is gas oil, but heavier diesel and wash oil streams also contribute to the total feed stream. Fresh feed is preheated in a heat exchanger and furnace and then passed to the riser, where it is mixed with hot, regenerated catalyst from the regenerator. Slurry from the main fractionator bottoms is also recycled to the riser. The hot catalyst provides the heat necessary for the endothermic cracking reactions. The gaseous cracked products are passed to the main fractionator for separation. Wet gas off the top of the main fractionator is elevated to the pressure of the lights end plant by the wet gas compressor. Further separation of light components occurs in this light end separation section.

As a result of the cracking process, a carbonaceous material, coke, is deposited on the surface of the catalyst, which depletes its catalytic property. Spent catalyst is recycled to the regenerator where it is mixed with air in a fluidized bed. Oxygen reacts with the deposited coke to produce carbon monoxide and carbon dioxide. Air is pumped to the regenerator with a high capacity combustion air blower and a smaller lift air blower. In addition to the contribution to the combustion process, air from the lift air blower assists the system with catalyst circulation. Complete details of the mechanistic simulation model for this particular model IV FCCU can be found in McFarlane et al. (1993).

Table 1 shows the selected predictor variables for the FCCU case study. All of these variables belong to the

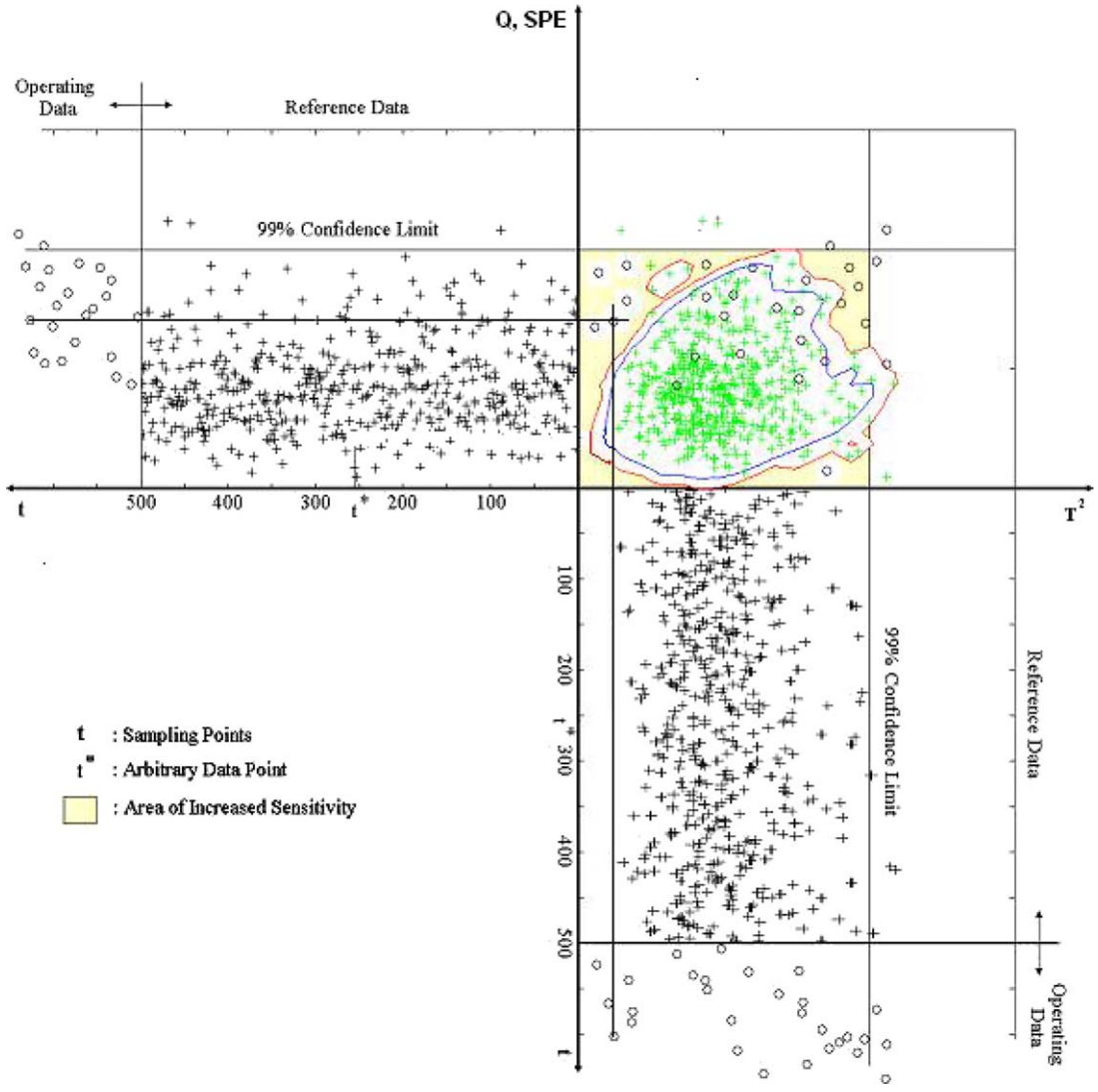

Fig. 1. Increased sensitivity of the joint analysis of the T^2 and Q statistics.

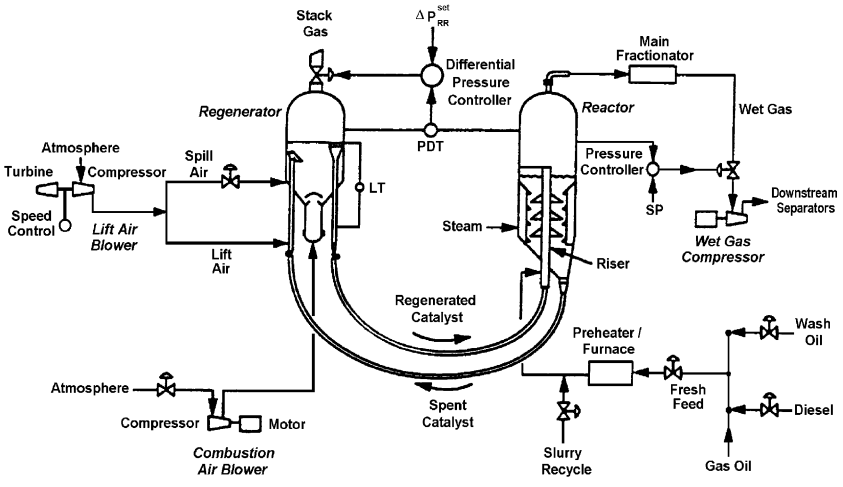

Fig. 2. Schematic diagram of the FCCU.

Table 1
Selected predictor variables for FCCU case study

Predictor variables	Signal
Diesel flow rate	ARMA sequence
Total fresh feed	ARIMA sequence
Slurry flow rate	ARIMA sequence
Preheat outlet temperature	ARIMA sequence

feed section of the unit. To simulate realistic disturbance conditions, various different autoregressive integrated moving average (ARIMA) signals were superimposed on these variables, with the exception of Diesel Flow rate, which received only an auto-regressive moving average (ARMA) signal, as indicated in Table 1. The response set includes Excess Oxygen in the Flue Gas, Concentration of Carbon Monoxide in the Flue Gas, Riser Temperature, Regenerator Bed Temperature, Regenerator Standpipe Level, Reactor Pressure as well as eight further measured variables from the system, see McFarlane et al. (1993) for a complete list of measured variables for the FCCU system.

Note that the application of ARIMA sequences violated the assumption of normally distributed variables. Hence, the non-parametric KDE technique was used to produce a more accurate PDF and consequently a more accurate confidence regions, which led to a more accurate statistical inference.

5.1. Description of the faulty condition

To test the performance of the presented approach, the FCCU simulator is augmented to include pre-programmed faults that could be applied on command. The considered fault was an incipiently developing increase (drift) in the friction for the flow of regenerated catalyst between the regenerator and riser, which is typically caused by partial or complete plugging of steam injectors located in this line. A reference data set, describing 2000 samples of normal operating condition, and a test data set, describing the abnormal condition, were sampled at a rate of one reading per minute. In the test data set, the drift was injected 440 points into the data set, equivalent to 440 min from the beginning of the recording. After 6200 samples of test data were recorded, the increased friction amounts to 0.1% of its normal level.

5.2. Result analysis

After identifying a PCA model using the reference data set, the samples of the T^2 and Q statistics were generated by applying Eqs. (4a) and (4b), including the 95% and 99% confidence limits for both statistics. These samples were then used to estimate the PDF using the KDE technique, as discussed in Sections 3 and 4. Fig. 3 displays the T^2 and Q statistics of the test data. Note that for each time-based plot, the abscissa represents time and the ordinate represents instances of a univariate statistic. Both, the T^2 and Q statistics, respectively given in Figs. 3(a) and (b), indicated that the drift could be detected after around 2850 samples into the data set, which means the average run length, i.e. the time span for detecting and abnormal event from when it arose, of this fault was to 2410 samples.

Fig. 4 shows the bivariate scatter diagram including the 95% and 99% confidence limits and the scatter points of the T^2 and Q statistics of the test data. It should be noted that the confidence regions and test data points spread into negative area although the T^2 and Q statistics can only produce values that are larger or equal to zero. This is related to the normalization and pre-whitening procedure, required for estimating the PDF using the KDE technique.

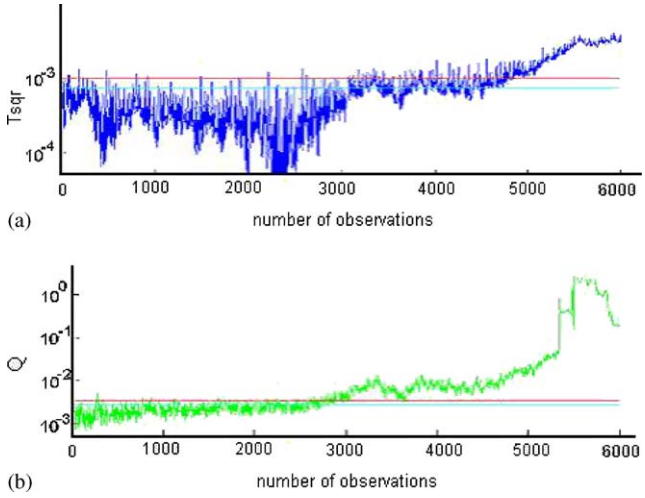

Fig. 3. PCA based T^2 and Q monitoring charts of FCCU application study.

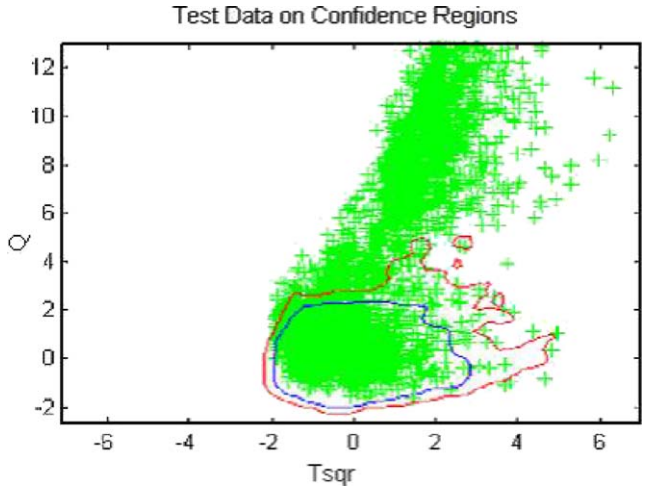

Fig. 4. Bivariate scatter diagram including 95% and 99% confidence regions of joint T^2 and Q analysis of FCCU system.

It can be seen from Fig. 4 that a vast number of scatter points that are outside the 99% confidence region. This picture also emerges when the univariate statistic, based on the density values of the joint PDF, is plotted with a time-based. Fig. 5 shows that this statistic leads to excessive violations of the 99% confidence limit after around 1640 min after the drift was injected, i.e. the average run length reduced to 1200 samples. This is, in fact, over 20 h earlier compared to the analysis of the individual T^2 and Q statistics. It can consequently be concluded that the joint analysis of the T^2 and Q statistics can be significantly more sensitive in detecting incipiently developing faults compared to the analysis of the individual statistics, as illustrated in Fig. 1.

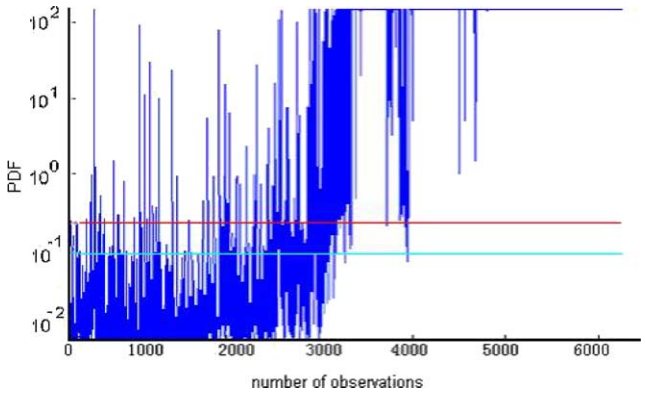

Fig. 5. Univariate monitoring chart on joint analysis of T^2 and Q and statistic for FCCU system.

6. A real process case—debutaniser distillation unit

The debutaniser distillation column, schematically shown in Fig. 6, separates butane (C_4) from a mixture of heavier hydrocarbons fed from an up-stream depolariser column. The Butane is taken from the top of the column and the heavier hydrocarbons are taken from the bottom of the column.

Large fan condensers condense the overhead stream and the liquid petroleum gas is fed to a reflux drum. The concentrations of propane (C_3) and hexane (C_5) in the reflux are measured by on-line analyzers and the reflux ratio is set accordingly to meet desirable impurity levels in the product stream. The bottom stream contains heavier hydrocarbons and is split between a feed back to the column, through a reboiler, and a feed to a down-stream processing operation. An on-line analyzer measures the concentration of C_4 in the bottom stream. It is desirable to minimize the C_4 concentration. Also, a sufficient level in the reboiler must be maintained to ensure that the steam coils are immersed and so avoid accretion of the coils. Process measurements taken from

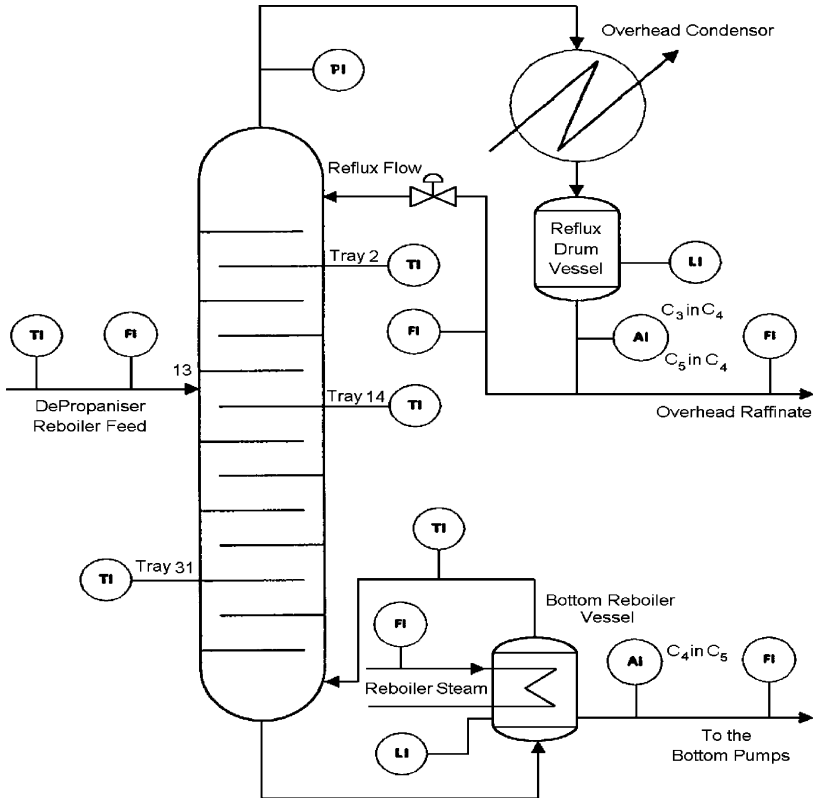

Fig. 6. Schematic diagram of the debutaniser distillation column.

Table 2
Debutaniser process variables

Type of variable	Notation
Controlled variable	Tray 14 temperature (°C)
Controlled variable	Column overhead pressure (bar)
Controlled variable	Tray 2 temperature (°C)
Controlled variable	Reflux vessel level (%)
Controlled variable	Butane product flow (T/h)
Controlled variable	Reflux flow (T/h)
Controlled variable	Analyzer of C_3 in C_4 (%)
Controlled variable	Analyzer of C_5 in C_4 (%)
Controlled variable	Tray 31 temperature (°C)
Controlled variable	Reboiler vessel level (%)
Controlled variable	Reboiler outlet flow (T/h)
Controlled variable	Analyzer of C_4 in C_5 (%)
Controlled variable	Debutaniser reboiler temperature (°C)
Manipulated variable	Reflux flow (T/h)
Manipulated variable	Reboiler steam flow (T/h)
Feed-forward variable	Depropaniser feed flow (T/h)
Feed-forward variable	Depropaniser reboiler temperature (°C)

the plant consist of 13 controlled variables, two manipulated variables and two feed-forward variables, as listed in Table 2.

6.1. Description of the fault condition

The investigated data contained undesired recurring increases of the C_5 fraction in the overhead C_4 product stream. This from time to time returning phenomenon could be traced back to a drop in the column feed that was sometimes accompanied by a slight drop in the feed temperature. This caused a reduction of the enthalpy stream entering the column, which almost instantly decreased the temperatures and pressures within the distillation column. The drop in fresh feed clearly caused the liquid level to become reduced, which by constant heat input of the reboiler gradually increased the outlet stream temperature of the reboiler. The enthalpy increase was reflected in the tray temperature sensors. When the temperature sensors exceeded specified thresholds, it could be noted that an impurity increase in the top draw followed inevitably and the concentration of C_3 in the bottom draw became reduced. From this process, two data sets were recorded that describe (i) normal process variation and (ii) the occurrence of a series of drops in fresh feed.

6.2. Result analysis

After the identification of a PLS model, retaining 3 LVs, and the generation of the T^2 and Q statistics, including their 95% and 99% confidence limits, the joint PDF of both statistics and the 95% and 99% confidence regions were estimated. The T^2 statistic, shown in Fig. 7(a) gave little indication of the presence of this event, whilst the Q statistic, given in Fig. 7(b), showed that around sampling instance 4100, frequent violations of the 99% confidence limit was noticeable and a less frequent but still statistically significant number of violations arose thereafter. Note that an insignificant drop in fresh feed arose after about 2500 and a series of significant drops emerged after around 4100 samples into the recorded data set. The impact of the latter event caused the discussed response of the unit and let to violations of the Q statistic beginning after around 4100 samples to the end of the recorded data set.

Figs. 8 and 9 show the bivariate scatter diagram and the PDF respectively. The statistic showed a similar picture to that of the PLS Q statistic, including a drop in fresh feed at around 2500 recorded samples was noticed by the joint analysis of both statistic. However, this drop

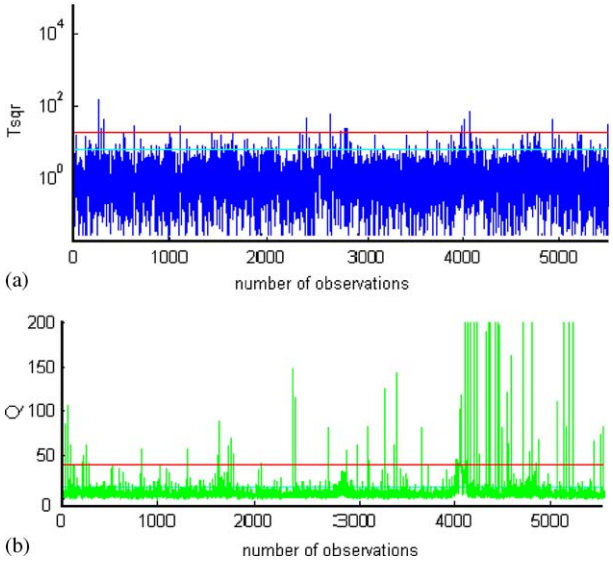

Fig. 7. PLS based T^2 and Q monitoring charts of industrial application study.

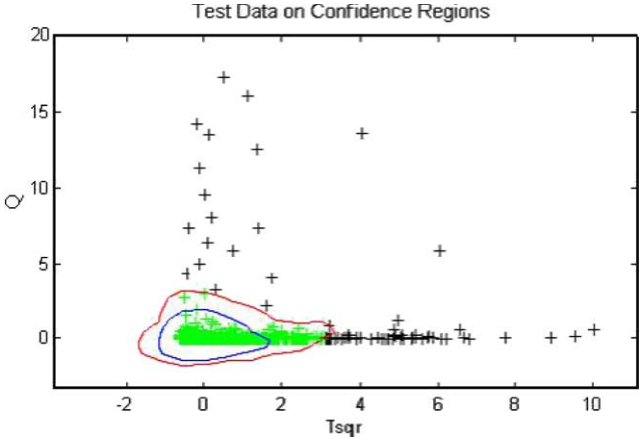

Fig. 8. Bivariate scatter diagram including 95% and 99% confidence regions of joint T^2 and Q analysis of industrial application study.

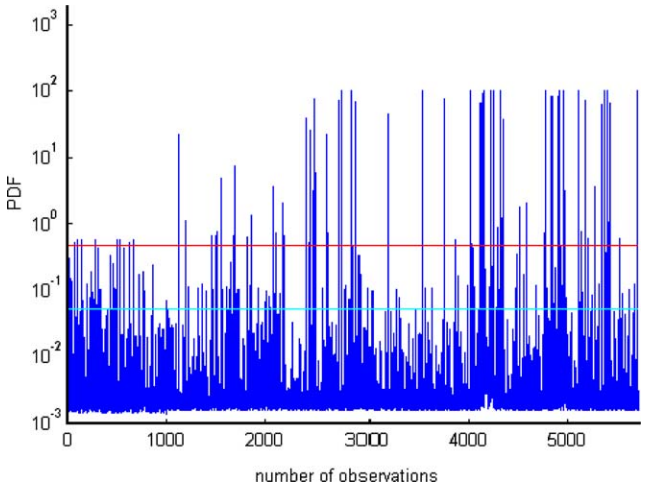

Fig. 9. Univariate monitoring chart on joint analysis of T^2 and Q and statistic for industrial application study.

was minor and went unnoticed by the PLS T^2 and Q statistics. This highlights again that the joint analysis of both statistics can be more sensitive than their individual analysis.

7. Conclusions

The novel approach of synthesis of the T^2 and Q statistics is a new way of using the T^2 and Q statistic data to monitor the operating performance of the process systems. The key advantage of the synthetic approach is the increased sensitivity as illustrated in Fig. 1 although it may not be necessarily shown in every case depending on the pattern of the faults. Also it is the further development of the application of KDE on process condition monitoring and fault detection. It is particularly useful in the case that the process distribution is unknown and T^2 and Q statistics are not chi-square distributed.

The main points of this approach are summarized as follows:

- *Speed-up*: With the introduced approach, all the problems caused by *curse of dimensionality* on PDF of more than two PCs or LVs can be avoided to speed up the process modelling procedure.
- *Simplification*: With the new approach, the number of monitoring charts has been reduced to one Shewhart chart of the joint PDF on $T^2 \sim Q$ or $T^2 \sim SPE$. For a clearer two-dimensional view, the corresponding contours of confidence regions of the joint PDF can be produced without any distortion.
- *Sensitivity*: With the new approach, the sensitivity of the monitoring charts to the abnormalities is improved.

Acknowledgements

The authors are grateful for the substantial and constructive comments by Dr. R.C. McFarlane of Invensys Performance Solution, Houston, Texas, USA concerning this article and the permission of accessing and helpful advice regarding the FCCU simulator. They acknowledge the assistance of Mr. S. Robinson of SimSci (UK) Ltd., in providing helpful advice in interpreting the process data of the distillation unit and gratefully acknowledge BP Exploration for providing the operating data used in this paper and for the permission to present associated results. This work is partially supported by the project funded by Natural Science Foundation Of China (grant number 60234010).

Appendix A

In order to show that the T^2 and the Q statistics for PCA and PLS are statistically independent it is required to investigate the process variables on which basis they are determined. The correlation between the residual of the process variables and the *s*-score variables are analyzed for PCA. This is followed by examining the correlation between the residuals of the response variables and the *s*-score vectors of the PLS algorithm.

A.1. Correlation between the residuals of the process variables and the s-score variables (PCA)

$$\begin{aligned}\mathbf{G}^{\mathrm{T}}\mathbf{S} &= [\mathbf{Z}-\hat{\mathbf{Z}}]^{\mathrm{T}}\mathbf{S}\\ &= \left[[\mathbf{S}\quad \tilde{\mathbf{S}}]\begin{bmatrix}\mathbf{R}^{\mathrm{T}}\\ \tilde{\mathbf{R}}^{\mathrm{T}}\end{bmatrix} - \mathbf{S}\mathbf{R}^{T}\right]^{\mathrm{T}}\mathbf{S}\\ &= \tilde{\mathbf{R}}\tilde{\mathbf{S}}^{\mathrm{T}}\mathbf{S} = 0,\end{aligned} \tag{A.1}$$

where $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{P}}$ represent the *s*-score and *r*-loading matrices, in which the discarded *s*-score and *r*-loading vectors are stored as columns in successive order. The above equation uses the fact that each of the *s*-score vectors is mutually orthogonal (Jackson, 1991).

Appendix B

B.1. Correlation between the residuals of the response variables and the t-score variables (PLS)

$$\begin{aligned}\mathbf{F}^{\mathrm{T}}\mathbf{T} &= [\mathbf{Y}-\hat{\mathbf{Y}}]^{\mathrm{T}}\mathbf{T}\\ &= [\mathbf{Y}-\mathbf{T}\mathbf{Q}^{\mathrm{T}}]^{\mathrm{T}}\mathbf{T}\\ &= [\mathbf{Y}-\mathbf{T}[\mathbf{T}^{\mathrm{T}}\mathbf{T}]^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y}]^{\mathrm{T}}\mathbf{T}\\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{T}-\mathbf{Y}^{\mathrm{T}}\mathbf{T} = 0.\end{aligned} \tag{B.1}$$

The above equation uses that the q-loading vectors are determined as follows:

$$\mathbf{q}_i = \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{t}_i}{\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i}. \tag{B.2}$$

B.2. Correlation between the residuals of the predictor variables and the t-score variables (PLS)

$$\begin{aligned}\mathbf{E}^{\mathrm{T}}\mathbf{T} &= [\mathbf{X} - \hat{\mathbf{X}}]^{\mathrm{T}}\mathbf{T} \\ &= [\mathbf{X} - \mathbf{T}\mathbf{P}^{\mathrm{T}}]^{\mathrm{T}}\mathbf{T} \\ &= [\mathbf{X} - \mathbf{T}[\mathbf{T}^{\mathrm{T}}\mathbf{T}]^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{X}]^{\mathrm{T}}\mathbf{T} \\ &= \mathbf{X}^{\mathrm{T}}\mathbf{T} - \mathbf{X}^{\mathrm{T}}\mathbf{T} = 0. \end{aligned} \tag{B.3}$$

The above equation uses that the q-loading vectors are determined as follows:

$$\mathbf{p}_i = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{t}_i}{\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i}. \tag{B.4}$$

B.3. Correlation between the residuals of the predictor variables and the response variables (PLS)

$$\begin{aligned}\mathbf{E}^{\mathrm{T}}\mathbf{F} &= [\mathbf{X} - \hat{\mathbf{X}}]^{\mathrm{T}}[\mathbf{Y} - \hat{\mathbf{Y}}] \\ &= [\mathbf{X} - \mathbf{T}\mathbf{P}^{\mathrm{T}}]^{\mathrm{T}}[\mathbf{Y} - \mathbf{T}\mathbf{Q}^{\mathrm{T}}] \\ &= \mathbf{X}^{\mathrm{T}}\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\mathbf{T}\mathbf{Q}^{\mathrm{T}} - \mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{Y} + \mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{Q}^{\mathrm{T}}. \end{aligned} \tag{B.5}$$

Since $\mathbf{X}^{\mathrm{T}}\mathbf{T} = \mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{T}$ and $\mathbf{T}^{\mathrm{T}}\mathbf{Y} = \mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{Q}^{\mathrm{T}}$, the above equation can be simplified as

$$\begin{aligned}\mathbf{E}^{\mathrm{T}}\mathbf{F} &= \mathbf{X}^{\mathrm{T}}\mathbf{Y} - \mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{Y} - \mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{Y} + \mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{Y} \\ &= \mathbf{X}^{\mathrm{T}}\mathbf{Y} - \mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{Y} = 0. \end{aligned} \tag{B.6}$$

If the $M - m$ discarded LVs are uninformative for predicting the response variables, the discarded t-score variables are uncorrelated to the response variables. This further implies that the covariance matrix, $[1/(K-1)]\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, can be reconstructed by incorporating the retained t-score vectors, i.e. the $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is equal to $\mathbf{P}\mathbf{T}^{\mathrm{T}}\mathbf{Y}$. Consequently, the residuals of the predictor and response variables are uncorrelated.

References

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, *71*, 353–360.

Chen, Q., Kruger, U., Meronk, M., & Leung, A. Y. T. (2001). Kernel density estimation on T^2 and Q-statistics for condition monitoring and fault detection. *Proceedings of the conference on condition monitoring*, Oxford, UK, 25–27 June.

Chen, Q., Wynne, R., Goulding, P., & Sandoz, D. J. (2000). The application of principal component analysis and kernel density estimation to enhance process monitoring. *Control Engineering Practice*, *8*(5), 531–543.

Chen, Q., Wynne, R., Sandoz, D. J., & Kruger, U. (1998). Improved kernel density estimation for clustered data using additive noise. *Proceedings of the American control conference*, Chicago, IL, USA, 28–30 June.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed). London: Academic Press, Inc.

Jackson, J. E. (1991). *A user's guide to principal components*. New York: Wiley.

Joliffe, I. T. (1986). *Principal component analysis*. New York: Springer.

Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, *28*, 3–21.

Kresta, J. V., MacGregor, J. F., & Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance. *The Canadian Journal of Chemical Engineering*, *69*, 35–47.

Kruger, U., Chen, Q., Sandoz, D. J., & McFarlane, R. C. (2001). Extended PLS approach for enhanced condition monitoring of industrial processes. *A.I.Ch.E. Journal*, *47*(9), 2076–2091.

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, *3*(3), 403–414.

MacGregor, J. F., Marlin, T. E., Kresta, J. V., & Skagerberg, B. (1991). Multivariate statistical methods in process analysis and control. *A.I.Ch.E Symposium proceedings of the 4th international conference on chemical process control*, *New York*: AIChE Publ. No. P-67 (pp. 79–99).

Martin, E. B., & Morris, A. J. (1996). Non-parametric confidence bounds for process performance monitoring charts. *Journal of Process Control*, *6*(6), 349–358.

McFarlane, R. C., Reineman, R. C., Bartee, J. F., & Georgakis, C. (1993). Dynamic Simulator for a model IV fluid catalytic cracking unit. *Computers and Chemical Engineering*, *17*(3), 275–300.

Morud, T. E. (1996). Multivariate statistical process control; example from the chemical process industry. *Journal of Chemometrics*, *10*, 669–675.

Nimmo, I. (1995). Adequately address abnormal situations operations. *Chemical Engineering Progress*, *91*(1), 36–45.

Silverman, B. W. (1986). *Density estimation*. London: Chapman.

Simonoff, J. S. (1996). *Smoothing methods in statistics, Springer series in statistics*. New York: Springer.

Valle, S., Weihua, L., & Qin, S. J. (1996). Selection of the number of principal components: The variance of reconstruction error criterion with a comparison to other methods. *Industrial and Engineering Chemical Research*, *38*, 4389–4401.

Vedam, H., & Venkatasubramanien, V. (1999). PCA-SDG based process monitoring and fault detection. *Control Engineering Practice*, *7*, 903–917.

Wand, M. P., & Jones, M. C. (1995a). Multivariate plug-in bandwidth selection. *Computational Statistics*, *9*, 97–116.

Wand, M. P., & Jones, M. C. (1995b). *Kernel smoothing*. London: Chapman.

Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, *6*(6), 329–348.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*, 37–52.