-

Automated Weighted Outlier Detection Technique for Multivariate Data

.

Suresh N. Thennadil[1*], Mark Dewar [2], Craig Herdsman[3], Alison Nordon[4], Edo Becker[3]

[1] *School of Engineering and IT, Charles Darwin University, Darwin, Australia*
[2]*Department of Chemical and Process Engineering, University of Strathclyde, Glasgow, UK*
[3]*BP Chemicals Ltd, Saltend, Hull, UK*
[4]*WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, UK*

**Corresponding Author: suresh.thennadil@cdu.edu.au (Fax: +61 8 8946 6680)*

ABSTRACT

In the chemical and petrochemical industries, spectroscopy-based online analysers are becoming common for process monitoring and control applications. A significant challenge in using these analysers as part of process monitoring and control loops is the large amount of personnel time required for calibration and maintenance of models which involve decision inputs such as whether an observation is an outlier, the number of latent variables in a model, type of pre-processing and when a calibration model has to be updated. Since no one measure works well for all applications, supervision by the process data analyst is required which invariably involves some level of subjectivity. In this paper, we focus on the detection of multivariate outliers in a calibration set. We propose a method which combines multiple outlier detection techniques to identify a set of outlying observations without operator input.

Apart from the overall methodology, this work introduces several novelties. The system uses partial least squares (PLS) instead of principal component analysis (PCA) which is normally used for detecting multivariate outliers. A simple modification to the Mahalanobis distance was also proposed which appears to be more sensitive to outliers than the conventional Mahalanobis distance. The methodology also introduces the concept of a desirability function to enable automatic decision making based on multiple statistical

measures for outlier detection. The methodology is demonstrated using Raman spectroscopy data collected from an industrial distillation process.

KEYWORDS

Multivariate outliers, Mahalanobis distance, Outlier detection, Desirability function, Multivariate Trimming.

Highlights

- An automated outlier detection system using multiple outlier measures weighted by a degree of anomaly function.
- A novel stopping criteria based on PLS regression model performance is proposed to choose the appropriate set of outliers.
- A simple modification to Mahalanobis distance measure is proposed and found to be more sensitive to outliers compared to the standard Mahalanobis distance.

1. INTRODUCTION

In the chemical and petrochemical industries, spectroscopy-based online analysers are becoming common for process monitoring and control applications [1, 2]. One of the challenges in the widespread usage of these technologies is the substantial amount of personnel time required for calibration and recalibration efforts. In addition, several steps in the calibration and maintenance of the models involve decision inputs such as deciding whether an observation is an outlier, the number of latent variables to be used in the model, the type of pre-processing and when to update the calibration models. These decisions are usually made in a subjective manner by personnel who use one or more statistical measures to aid them in their decision. Thus, this type of work also needs personnel with a high level of expertise and skills in building and maintaining calibration models. It is therefore highly desirable to have an automated system for carrying out these tasks since such a system can deliver significant cost savings whilst also providing a consistent approach to modelling.

While several statistical measures are available for objectively making the necessary decisions, no single measure has been shown to consistently perform for outlier detection or for choosing the parameters of a calibration model.

An outlier can be defined as an "observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [3]. Outliers present in the data can be the result of any number of things including an instrument fault, a process disturbance and even instrument drift. The presence of outliers in a dataset can dramatically undermine the analysis and any subsequent results based on the data. When dealing with spectroscopic data, which consist of measurements at several wavelengths, the possibility of the occurrence of multivariate outliers has to be considered. In other words, the spectrum as a whole has to be evaluated to decide if an observation (spectrum) is an outlier.

Two of the most frequently occurring deleterious effects of outliers (both univariate and multivariate) are that of masking and swamping. Masking is the effect of outliers, or clusters of outliers, present in the data skewing the mean and covariance towards themselves. This can reduce the distances of the outlying observations from the mean, resulting in some of them being placed within the region enclosed by the confidence limits and thus be mistakenly considered as regular observations. Thus, some of the more extreme outliers which are identified even with the skewed mean and covariance can effectively mask the outlier status of other less extreme outlying observations. Once the identified outliers are removed the 'mask' is removed since the new values of mean and covariance will lead to a narrower confidence region and those outliers that were previously misclassified as normal observations may be revealed as outliers [4]. Swamping is the effect caused by outliers, or clusters of outliers, present in the data, skewing the mean and covariance away from non-outlier observations. This causes the non-outlier samples to have a relatively large distance from the mean and hence have the artificial appearance of an outlier in the dataset. An outlier is said to swamp another observation when in the presence of outliers, a normal observation appears to be an outlier [5].

To understand and correct for the presence of outliers in a dataset, identifying their presence in the dataset is the first stage. At present the process of identifying an observation as being an outlier is, for the most part, a largely subjective process. It can also be a considerably time-consuming practice considering the many different possible outlier detection criteria available and especially when multiple or large datasets are required to be analysed. It is therefore desirable to have a system that is able to quickly and objectively identify, and potentially remove, the samples that are branded as outliers without the requirement of subjective decision making and specialist knowledge.

Numerous outlier detection methods for multivariate data (which is the focus of this paper) currently exist. The shortcoming of several of them is that they are not able to robustly and accurately detect outliers for all possible datasets. Many of the methods that currently exist rely on distance-based measures (for example, Hotelling's T^2 and Mahalanobis distance) and visualisation techniques.

Lu et al.[6] proposed an advancement in the detection of univariate outliers through the use of the median in place of the mean in a spatial outlier detection algorithm pointing out that the median is a more robust estimator of the "centre" of a dataset. Shekhar et al.[7] offered a technique using graph structured datasets through which an attribute value and the average attribute value of its neighbours provide a distinction for identifying outliers significantly different from those of their neighbourhood. Wilson wrote on a statistical methodology for detecting outliers by ranking observations in order of their dissimilarity to the others in the dataset[8]. Though all these techniques have their merit in detecting outliers, there is difficulty in establishing any kind of superiority between the methods. This is because a direct comparison between outlier detection methods is not always possible as the efficiency of different methods depends on different criteria, such as the dimension of the data set, the type of the outliers, the proportion of outliers in the dataset, and the outliers' degree of anomaly.

In a 2001 paper, Penny and Joliffe [9] concluded that due to the difficulty with direct comparison and the varying degrees of accuracy for any given outlier test when applied to different datasets, the best option would be to use various outlier tests together in a "battery" combination. This work applies their idea by constructing a battery of multivariate outlier detection methods in a way that is part of an automatic and objective system.

To achieve objectivity in the detection of outliers, a methodology is proposed whereby all samples being analysed are subjected to a battery of outlier detection tests. If a sample fails an individual test that sample gains an outlier weighting dependent on the samples' relative

degree of anomaly and depending on the particular outlier test failed. This would mean that an observation, which falls beyond a set confidence limit for a particular test, will receive a weighting relative to its distance from the limit. If an observation fails multiple tests, its outlier weighting accumulates according to the weightings associated with the particular tests that it failed.

There are essentially two ways to approach the detection and removal of outliers from a dataset. One approach is to identify the outliers in one step where a statistical test is used to identify the outlying observations which are then removed from the dataset. This approach has a potential pitfall. Since the statistical measure which is usually some kind of distance measure using the mean and covariance of the dataset can be influenced by the outliers, masking and swamping effects can lead to erroneous classification of the observations into outlying or regular observations. Though the system applies multiple outlier detection tests in combination, it does not use a single-step analysis of the outliers, whereby the detection of outliers is undertaken in one step, as is suggested by Davies and Gather [10]. Their proposed method, which relies on robust estimation of the sample mean and sample standard deviation, can therefore be significantly skewed in the presence of outliers.

In the methodology proposed in this paper, once all observations are given an associated outlier weighting based on their passing or failing the various tests, each sample is given a ranking that corresponds with its degree of anomaly (which is informed by their individual outlier weightings). Through this ranking structure, a single observation that deviates the most from the rest of the data is removed and the model is recalibrated and the process repeated. This method for removal of the most established outlier forms the basis of the multivariate trimming process used in the proposed methodology. The potential advantage of using this multivariate trimming approach over conventional methods (where the detected outliers are compiled together in a reservoir and then either deleted or returned to the main body of the

data based on whether or not they are considered an outlier) is that through the proposed approach, the data is not as susceptible to the possible skewing effects of outliers [11].

In this paper, the focus will be on developing an automated system for outlier detection. In particular, an approach will be described for the case where a calibration dataset is available and the methodology is used to detect and remove outliers from the dataset as part of the model building process. The goal of the proposed approach is to provide a methodology that will pick more or less the same set of observations as outliers that would be picked by a human expert. The rest of the paper will thus deal specifically with outlier detection. The proposed approach has the advantage of being extendable into an integrated automated calibration model building approach. Further, the proposed outlier detection approach can be easily modified to be applicable for detecting outlying measurements during online process monitoring.

In the next section, the outlier tests used as part of the methodology will be described. This will be followed by a description of the methodology which combines these tests to decide whether an observation is an outlier. The individual tests and the proposed method of weighted outlier detection test will be demonstrated by applying them to a dataset which consists of Raman spectroscopy measurements taken from a petrochemical distillation unit.

2. THEORY

The methodology as described in this paper uses 4 outlier detection tests. Three of these tests are based on the spectral measurements (X-block), namely scores confidence limits, modified Mahalanobis distance and a test based on the Q residuals and Hotelling's T^2 which

will be referred to as the leverage test, while one is based on the property of interest (Y-block) using the Y-residual value.

2.1 Scores Confidence Limit Test

Confidence limits have been commonly used for identifying univariate and multivariate outliers. In the latter case, the spectra have to be first transformed into the latent variables (LVs) domain. Usually principal component analysis (PCA) is used for this purpose [12]. The number of latent variables required to explain the majority (usually >90%) of the variation in the dataset is chosen to build the PCA model. Each observation (spectrum) is then decomposed using PCA to provide the scores associated with each LV. Based on the scores obtained for the entire dataset, confidence intervals are calculated for the scores of each latent variable. This is essentially the application of the detection technique for univariate outliers since we are examining one LV at a time. For a particular spectrum, the score for each latent variable is considered and if the score of any of the latent variables falls beyond the confidence interval, the spectrum is flagged as an outlier. In practice, most of the variation in the data is captured in the first 2-3 latent variables and it can be expected that the effect of an outlier will mostly manifest in the first few LVs. While PCA is commonly used for this type of outlier detection, in this study, scores from partial least squares regression (PLSR) are used. Since the goal is to remove outlying observations that would have an adverse impact on calibration models built with PLS, it is more logical to use the same decomposition method for detecting outliers.

Assuming that the scores for a latent variable follow a normal distribution, the confidence interval for the scores can be calculated using the T-statistic which follows the Student's T-distribution. The confidence interval is given by [13]:

$$x_{ij} - t_{\alpha/2,n-1} s_j \leq \mu \leq x_{ij} + t_{\alpha/2,n-1} s_j \quad (1)$$

where x_{ij} is the score of the i[th] observation for the j[th] latent variable and s_j is the sample standard deviation for the j[th] latent variable. The n – 1 subscript in the formula refers to the degrees of freedom with n being the number of spectra in the dataset and α is the probability that an observation (spectrum) will fall outside the confidence interval. It is related to the confidence level (CL) which is calculated as follows:

$$\alpha = 1 - \frac{CL}{100} \quad (2)$$

Spectra with scores values x_{ij} that fall outside the confidence interval are considered to be outliers. In the method proposed here, the traditional confidence limit test is modified to combine an absolute limit with a weighting function. A measure of the "degree of anomaly" is calculated based on an observation's distance from the confidence interval. This approach consists of setting two confidence intervals: one at the 95% confidence level (α = 0.05) and another at the 99% confidence level (α = 0.01). If the score of the observation falls outside the first interval (95%), it is considered as an outlier. Subsequently the degree of anomaly (w_i) of the identified outlier is calculated using the equation given below.

$$w_i = \begin{cases} \dfrac{x_{ij} - x_{95\%_U}}{\left(x_{99\%_U} - x_{95\%_U}\right)} & if\ x_{ij} > 0 \\ \dfrac{x_{ij} - x_{95\%_L}}{\left(x_{99\%_L} - x_{95\%_L}\right)} & if\ x_{ij} < 0 \end{cases} \quad (3)$$

In equation (3), w_i is the degree of anomaly of spectrum *i*, x_{ij} is the scores value of the observation *i* for latent variable *j* and $x_{95\%}$ and $x_{99\%}$ are the critical values for the confidence limits of 95% and 99% confidence levels respectively. The subscripts *U* and *L* indicate the upper and lower confidence limit respectively. The weighting is calculated as the relative

difference between the two limits and the test value of the observation. The weighting becomes greater than 1 if an observation is over the 99% limit.

Calculating the level of deviation from the 95% confidence interval in this way allows for a relative and continuous weighting to be applied to the outlier detection test. This has the advantage over a discrete form of weighting as it reduces the probability for the occasion where two or more samples are given the exact same weighting but have different degrees of anomaly (as would be the case if there were only one set of confidence intervals beyond which all samples received the same weighting). Due to the weighting being based on relative difference, weightings from different tests will be comparable.

2.2 Mahalanobis Distance

Distance-based measures can be used to automatically locate multivariate observations which are far from the centre of the dataset. The Mahalanobis distance (MD) [14] is a measure of distance of each observation while accounting for correlations between variables as well as the differences in variances between those variables[15]. The Mahalanobis distance (MD_i) for each multivariate observation (spectrum) i is given by:

$$MD_i = [(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})]^{1/2} \quad (4)$$

where $\boldsymbol{x}_i$is a column vector containing readings (e.g. absorbance) at different wavelengths of the spectrum i, $\overline{\boldsymbol{x}_i}$ is the mean spectrum of the dataset and $\boldsymbol{S}$ is the variance-covariance matrix. This is the classical Mahalanobis distance formula as used by Penny and Jolliffe in their study of multivariate outlier detection tests applied to medical data [16].The squared Mahalanobis distance follows a Chi-squared distribution [17]. While in principle, (4) can be applied to the spectral dataset, the high correlation between readings at different wavelengths results in S being too ill-conditioned for inversion. This problem can be circumvented by first

transforming the spectral data matrix to an orthogonal scores matrix through decomposition using PCA or PLS. In this case, (4) is used by substituting the raw measurements of spectrum *i* by the corresponding scores vector and ***S*** with a diagonal matrix of eigenvalues of the latent variables. When using scores instead of the raw measurements, the number of latent variables is chosen so that close to 100% of the variance in the spectral data is included. For practical purposes, in this study, the number of latent variables were selected so that >99.9% of the variance in the X-block or the Y-block (whichever is achieved first) was explained. Those observations with a significantly large Mahalanobis distance are indicated as outliers.

Confidence intervals for the squared Mahalanobis distance have been constructed in different ways. One approach is to use the chi-squared statistic. In this case it is assumed that the sample size is large enough so that the computed sample mean and covariance matrix are very close to the population (expected) values. In other words, the chi-squared distribution is achieved in an asymptotic sense. This approach has been used by Shah and Gemperline in their analysis of NIR spectra for classifying materials [18] and more recently by Liu and Weng, for analysing satellite image data [19]. Confidence intervals based on the F distribution (or the equivalent T^2 distribution) have been used by a number of researchers to account for the fact that the sample mean and covariance matrix are used [20,21,22]. Vervaridis and Kotropoulos [23] consider the case where the sample mean and covariance matrix are estimated by including the possible outlying points. They showed that the Mahalanobis distance would then follow the beta distribution.

In this study, we have used the chi-squared statistic for calculating the confidence limits. The confidence limit for a given confidence level M_{CL} is given by:

$$M_{CL} = \sqrt{\chi^2_{(1-\alpha/2,n-1)}} \qquad (5)$$

where χ^2 is the Chi-squared value with (n-1) degrees of freedom and probability α. As with the previous test, two confidence limits are set at 95% and 99% giving a boundary and a weighting similar to (3) is used for identifying and weighting the outliers:

$$w_i = \frac{M_i - M_{95\%}}{(M_{99\%} - M_{95\%})} \tag{6}$$

2.3 Modified Mahalanobis Distance

The classical Mahalanobis distance is limited in its ability to detect outliers as the formula itself is susceptible to the effects of outliers. Specifically the classical Mahalanobis distance is susceptible to the effects of masking[24] and swamping. This is mainly due to the reliance of the classical Mahalanobis distance on the mean and variance/covariance which can be skewed by outliers.

Rousseeuw[25] proposed a method to calculate a robust Mahalanobis distance using the minimum covariance determinant (MCD). This method has been investigated by several groups either in the original or in modified forms [22,26,27]. This method is iterative and computationally intensive.

A simple "Robust" Mahalanobis distance measure, which is a modified form of the classical Mahalanobis distance, is proposed in this study and consists of replacing the mean and covariance with their robust counterparts namely, the median and the interquartile range, respectively. The modified Mahalanobis distance (MD_{iR}) is then given by,

$$MD_{iR} = \left[(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_m)^T \boldsymbol{S}_{iqr}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_m)\right]^{1/2} \tag{7}$$

where $\overline{\boldsymbol{x}}_m$ is the median vector of score values from PCA or PLS and $\boldsymbol{S}_{iqr}$ represents the square of the interquartile range which is the statistical dispersion measured as the upper quartile minus the lower [28]. The $\boldsymbol{S}_{iqr}$ is a robust estimate of the spread of the data, since

changes in the upper and lower 25% of the data do not affect it. It is assumed that the use of median and IQR in place of the mean and covariance does not affect the distribution of the Mahalanobis distance. Though it has not previously been applied to the Mahalanobis distance the application of the median in place of the mean in other outlier detection tests, as a more robust alternative, was examined by Lu et al [6]. In their study, three different spatial outlier detection tests were compared, with one of them using the median in preference to the mean. The results showed that the median based robust test was more effective than the other tests in the detection of outliers.

2.4 Leverage Outlier Test

The leverage test is made up of two separate confidence limits based tests. The first test is that which is based on Hotelling's T^2 distribution test [29]. The values of T^2 give an indication of the samples' distance from the centre (multivariate mean) of the model [30]. The use of Hotelling's T^2 statistic for the identification of outliers in a multivariate system is a popular method which has also been used in the construction of control charts [31]. T^2 is the sum of the normalised squared scores and is defined by:

$$T_i^2 = \boldsymbol{t_i \Lambda^{-1} t_i^T} \qquad (8)$$

where T_i^2 is the Hotelling T^2 value for spectrum *i*, $\boldsymbol{t_i}$ is the scores vector for the spectrum and $\boldsymbol{\Lambda}$ is the diagonal matrix consisting of the eigenvalues of (or variance explained by) the latent variables included in the model.

It should be noted that the squared Mahalanobis distance is similar to the Hotelling's T^2, when the scores are used instead of the raw measurements to calculate (4), and (8) is calculated using scores derived from mean-centred spectra and the model explains 100% of the variance in the data. This will occur if the number of latent variables included in the model

corresponds to the dimensionality of the raw data. In this study, the robust approaches for the calculation of Hotelling's T^2 are not considered. Instead, the outlier detection is carried out by using a combination of T^2 and the Q residuals.

As indicated above, the Hotelling T^2 is similar to the Mahalanobis distance and thus a robust estimate could be obtained using the minimum covariance determinant (MCD). Such an approach was used by Vargas[32] and Jensen et al.[33]. However, as stated by Shabbak et al.[34], their work is only evaluated based on the number of outliers detected and not whether or not those detected observations were indeed true outliers.

The second test, which forms the other half of the leverage outlier detection test, is based on the Q residual (in spectroscopy applications this is commonly referred to as spectral residuals). Q residuals are a lack of fit statistic. The values for Q residuals denote the amount of the variation which remains in each spectrum after projection through the model. Q residuals are calculated through the sum of squares of the residual error in the spectra reconstructed using the model compared to the measured spectra. The Q residual is therefore capable of contributing to the determination as to whether or not any of the lack-of-fit present in the model is the result of random variation or the presence of systematic variation. The Q residual for spectrum *i* is given by,

$$Q_i = \boldsymbol{e}_i^T \boldsymbol{e}_i \qquad (9)$$

where $\boldsymbol{e}_i = \hat{\boldsymbol{x}}_i - \boldsymbol{x}_i$is error i.e. the difference between the measured spectrum (x_i) and corresponding spectrum calculated using the model ($\hat{x}_i$).

The leverage outlier detection test combines these two sub-tests. A spectrum that fails only one of the subtest is not considered to be an outlier. If the observation fails both tests, then it is considered an outlier.

The confidence limits required for the making the decision are given below. The limit for the Hotelling T^2 is given by[35]

$$T^2_{K,n,\alpha} = \frac{K(n-1)(n+1)}{n(n-K)} F_{K,n-K,\alpha} \tag{10}$$

where K denotes the number of latent variables used by the model, *n* is the number of spectra in the data set, $F_{K,n-K,\alpha}$ is the critical value of the F distribution with K, *n* - *K* degrees of freedom with confidence level associated with probability α as given in (2).

The limit for the Q residual is given by[36]

$$Q_\alpha = \Theta_1 \left[\frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{{\Theta_1}^2} \right]^{\frac{1}{h_0}} \tag{11}$$

where

$$h_0 = 1 - \frac{2\Theta_1 \Theta_3}{3\Theta_2^2} \tag{12}$$

and

$$\Theta_i = \sum_{j=K+1}^{M} \lambda_j^i \ \ for\ i = 1,2,3 \tag{13}$$

where c_a is the critical value of a standard normal random distribution for probability α, M is the maximum number of latent variables i.e. the rank of the spectral matrix and λ_j is the eigenvalue associated with latent variable *j* [30]. For an observation that fails both sub-tests, the degree of anomaly given by,

$$w_i = 0.5 \left[\frac{Q_i - Q_{95\%}}{(Q_{99\%} - Q_{95\%})} + \frac{T_i^2 - T_{95\%}^2}{(T_{99\%}^2 - T_{95\%}^2)} \right] \tag{14}$$

2.5 Y Residual Test

The Y residual test is used to identify outliers in data relating to the dependent variable (property of interest). The error in the estimation of the property value corresponding to spectrum i is the difference between the actual (experimentally determined) value y_i and the predicted value $\hat{y}_i$ i.e. $\varepsilon_i = \hat{y}_i - y_i$. This error difference is used as the basis for this test. It should be noted that this particular test can only be used at the calibration stage of modelling. This is due to the need for reference values of *y*. The confidence interval corresponding to probability α for the error in the estimate of y is given by

$$\varepsilon_\alpha = \pm t_{\frac{\alpha}{2},n-1} \cdot s \qquad (15)$$

where *s* is the root mean square error. As in the case of the scores confidence test, a weighting for the anomaly of the sample is calculated using an equivalent of equation (3). This test is related to the Predicted Error Sum of Squares (PRESS) statistic that is used for the most part for making an assessment of the quality in the model predicted but can also be applied to outlier detection[37].

2.6 Desirability Function

Occasionally the data used in a particular analysis may be more sensitive to certain tests than others or based on experience it may be found that certain tests are more reliable for specific data types[9] . In such cases it is advantageous to use an additional weighting for the individual outlier tests in a manner that reflects the effectiveness of that test. One approach to introduce such weighting is the use of a desirability function. In addition to the continuous weightings that are associated with each of the outlier detection tests, there is an overall weighting used as a measure for the significance of each test.

The desirability function approach has been used in multiple output optimisations[38,39] which are able to optimise for multiple responses. A comparative study conducted by Chakravorty et al. [40] was done on the effectiveness of several multi-response optimisation techniques. It was concluded that the use of a desirability function was the strongest option for optimisation. Although the study focused on the use of these optimisation techniques with ultrasonic machining processes the results may be translated to the work carried out in this paper. A further study by Costa et al. [41] compared several different desirability functions. Included in this comparison was the function by Derringer and Suich.[39] It was concluded that no one method necessarily outperformed another. However, due to certain qualities of the function proposed by Ch'ng et al. [42] it was identified as potentially being slightly more advantageous. A desirability function of the form given below is used to calculate the global desirability (D),

$$D = (w_1^{p_1} + w_2^{p_2} + \cdots + w_n^{p_n}) \qquad (16)$$

where w_1, w_2,…. are the individual desirability (or undesirability in the case of outliers) and p_1, p_2…. are the associated significance parameters. In the context of the weighted outlier detection test, w_i will be the degree of anomaly calculated for outlier test i and p_i will be the level of importance (expressed as a fraction) associated with that test. An example would be supposing that Mahalanobis Distance was considered as being of high importance to the detection of outliers, it would have a high percentage weighting say 60% (p = 0.6). It should be noted that as defined in equations 3 and 6, w_i will always be greater than one. By utilising prior knowledge, if available, a judicial choice of p_i can be used to deliver a more reliable outcome in terms of identifying the outliers. For this study, the tests were assigned equal percentage significance since prior knowledge in terms of the effectiveness of outlier test for the type of data sets considered did not exist.

The methodology described below is not restricted to the use of the 4 outlier detection methods described in this section. Any number and combination of outlier detection methods can be used. The four methods used in this study include methods commonly used with added novelties such as modification to the Mahalanobis distance (robust version), the introduction of measures for weighting the degree of anomaly and the desirability index.

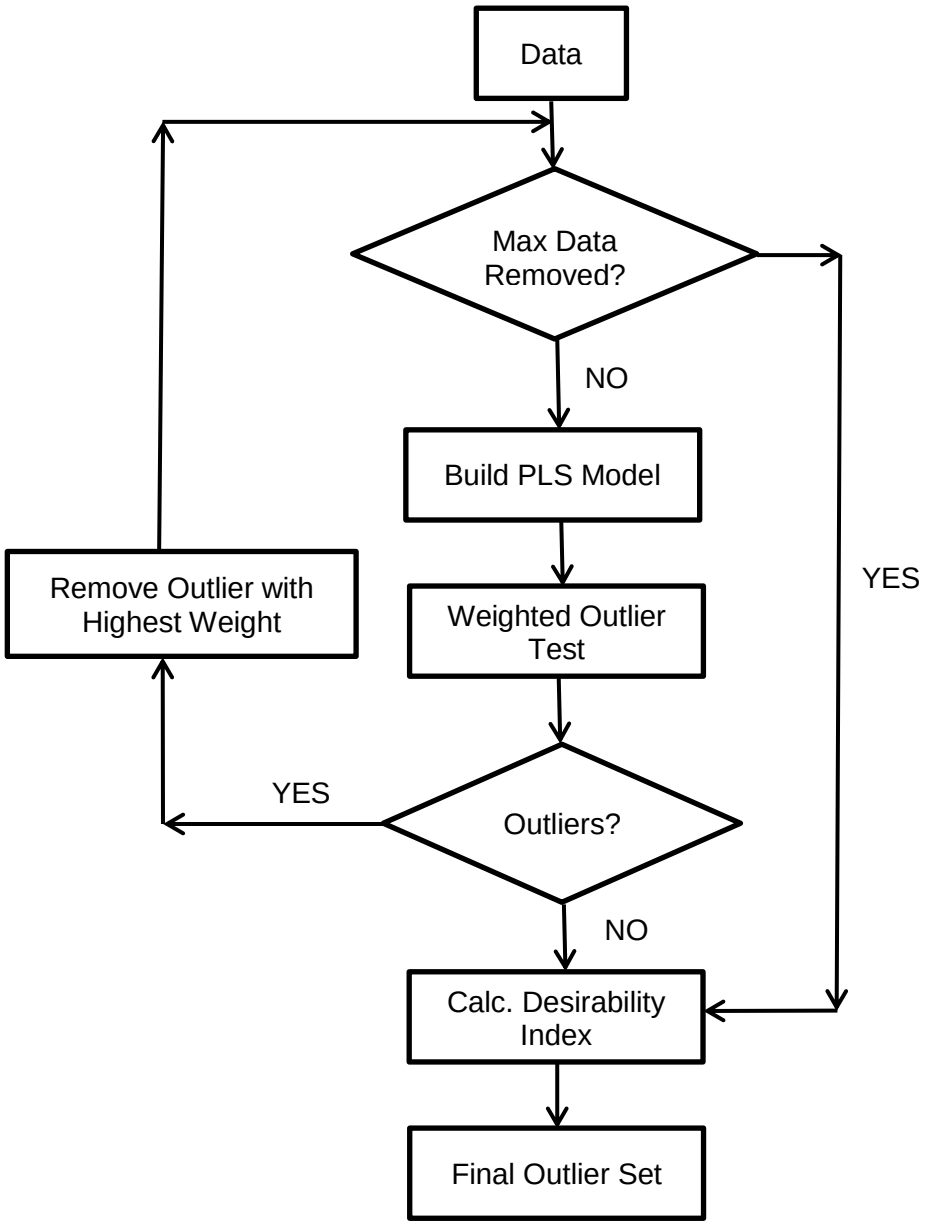

Figure 1: Flowchart of the procedure used by the proposed system to effectively detect outliers.

3. METHODOLOGY

Figure 1 shows the flowchart illustrating the application of the proposed automatic integrated outlier detection method. The first step is to build a PLS model. This can be preceded by appropriate pre-processing. However, in this study, no pre-processing of the spectra other than mean-centring was performed. The number of latent variables to be chosen is based on cross-validation. Many different methods are available to automatically estimate the optimal number of latent variables. In this paper the optimal number of factors is automatically selected using Wold's R criterion[43]:

$$R = \frac{PRESS(a+1)}{PRESS(a)} \tag{17}$$

The optimum number of LVs was taken to be the number beyond which R > 0.9[44]. Once the optimum number of latent variables is chosen based on cross-validation, the scores and loadings for this model are calculated. As mentioned earlier, normally PCA is used to identify X-block outliers. Here PLS is used since it can naturally be applied to identification of the y-outliers. The impact of using PLS instead of PCA is discussed in the results and discussion section.

Once the model is built using PLS, the data is then subjected to a series of outlier detection tests which are described in the Theory section. A spectrum that fails any of the individual tests will receive a weighting according to the degree of anomaly calculated using the appropriate weight function. The overall degree of anomaly (D) for an outlying observation is then calculated using the desirability index given by (16). It should be noted that values of p_1, p_2… are based on *a-priori* knowledge of the reliability of each outlier detection technique which will have to be based on experience. In this study, since at this point such *a-priori* knowledge is not available, the tests are assigned equal significance (i.e. p_i values of 1) thus

reducing the equation to a sum of the degree of anomaly returned by each of the individual tests.

To minimise the chance of masking and swamping effects causing incorrect identification of outliers, a form of multivariate trimming (MVT) is used [45, 46]. While several observations can be identified as outliers, only a fraction of them are removed. Using this concept, the observations with the k largest values of D where k is a predetermined number are removed. In this study, a conservative approach was used whereby only the spectrum with the highest overall degree of anomaly *D* (i.e. k = 1) calculated using (16) is removed from the data set. The remainder of the data is then used to develop a new PLS model. The outlier tests are then carried out again to identify the observation with the largest overall degree of anomaly. The process of building PLS models using the remaining data and identifying a set of outliers is repeated until no new outliers are identified or if the number of spectra removed from the dataset exceeds a set percentage. The latter limit is introduced since the MVT process can, in many cases, continue without converging i.e. without reaching a stage where no new outliers are found. This can result in increasing chance for observations to be wrongly identified as outliers due to the decrease in the sample size which will affect the accuracy of the scores and loadings calculated using PLS as well as the increase in uncertainties in the parameters such as standard deviation, Mahalanobis distance etc. If the loop is exited because the maximum allowable number of spectra has been removed, then a final PLS model using the remaining data is built to generate the model performance data required for the next step. If instead the loop was terminated due to no new outliers being found, then the PLS model built prior to the outlier detection step will be the final model since the dataset will not have changed after the detection step.

An additional step is used to combat the issue of removing spectra which are not true outliers. It should be noted that in this study, the outliers are considered ultimately from the

point of their adverse effect on model performance. The removal of a set of observations which are true outliers can be expected to improve model performance. If they are not actual outliers, then the impact of having fewer spectra for calibration can lead to increased uncertainty in the model parameters which can lead to a possible degradation in model performance. Thus, it is logical to only remove the set of observations whose removal will result in a positive impact on model performance. For this purpose, the following desirability index is proposed:

$$D_i = \left(\frac{y_{RMS,min}}{y_{RMS,i}} + \frac{F_i}{F_{i,max}}\right) * (1 - \gamma \frac{i}{I_{Trim}}) \quad (18)$$

where the subscript i refers to the ith trimming step, y_{RMS} is the root mean square error of cross-validation (RMSECV), $y_{RMS,min}$ is the minimum value of RMSECV amongst the I_{Trim} values generated and I_{trim} is the number of trimming steps at which further outliers are not observed. This will have a maximum value which is based on the percentage of data that can be trimmed. In this study, this was set to 20% of the total number of observations which translates to 19 trimming steps for the dataset considered in this study. F is the F-ratio which is given by[47]:

$$F_i = \frac{MS_R}{s^2} \quad (19)$$

MS_R is the mean-square due to regression and s^2 is the mean-square due to residuals. $F_{i,max}$ is the maximum value of F obtained over the I_{trim} steps. The term $(1 - \gamma \frac{i}{I_{Trim}})$ is a penalty term introduced to account for the effect of the reduced number of observations on the uncertainties in estimates of scores. It can also account for the possibility of variations that belong to the population being removed and thus have a negative impact on the future performance of the model. The parameter $\gamma \leq 1$ is used to adjust the sensitivity of the desirability index to the

number of outliers removed. It should be noted that other measures such as the number of latent variables and R^2 can be included as part of the desirability index.

4. **DATASET**

The dataset consists of Raman spectra collected from a petrochemical distillation column. The column is used to remove cyclic and longer chain hydrocarbons from hydrocarbon feed which contains paraffins, aromatics and naphthenes. Raman spectra were collected using a Kaiser Holoprobe spectrometer with a remote laser assembly on a translation stage. Each spectrum consisted of intensities at Raman shifts spanning 0 – 1892 cm^{-1} at wavenumber intervals of 1 cm^{-1}. Reference values for the concentrations of the hydrocarbon of interest, which will be referred to as Component A in this paper, were obtained from samples corresponding to the spectral measurements, using a standard reference method followed by BP. The dataset consisted of 99 observations encompassing a wide range of process variations. Outliers in the spectral data were identified by personnel in BP using visualisation, knowledge of process disturbances and by standard PCA. The outliers decided by analysis conducted by the personnel are taken as "known" outliers against which the automated system will be compared in order to evaluate whether the automated methodology will pick the same set of observations as outliers. The dataset was found to have 8 "known" outliers namely observations: 1, 2, 3, 39, 40, 41, 42, and 43. The Raman spectra for the 99 observations are shown in Figure 2.

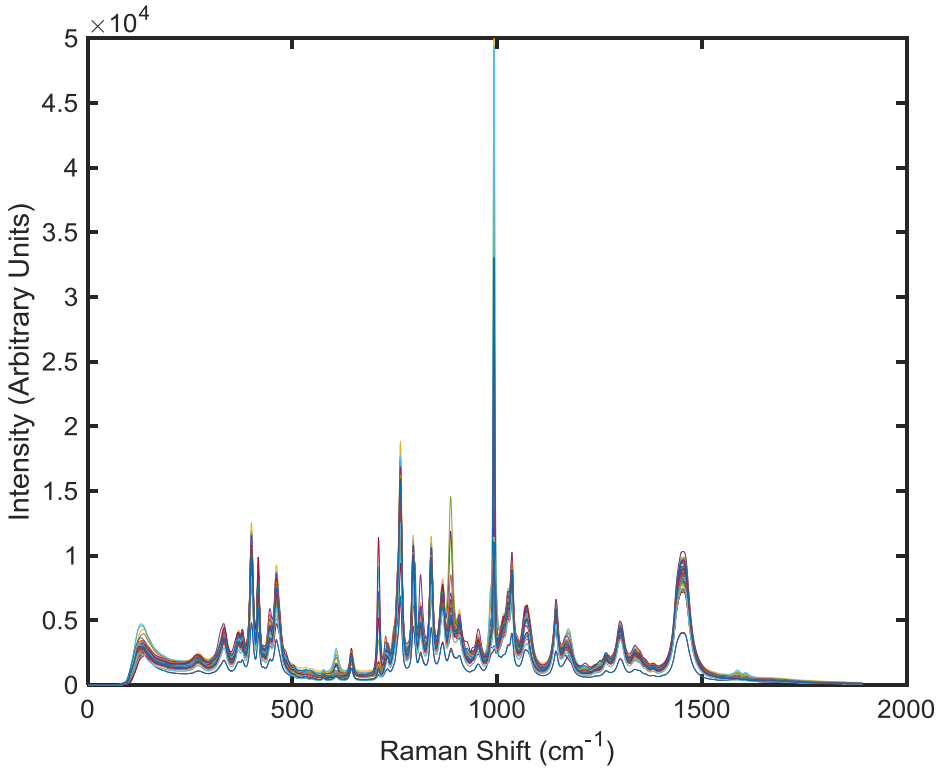

Figure 2. Raman spectra of the light hydrocarbon stream.

5. RESULTS & DISCUSSION

As previously mentioned, we use PLS instead of the usual PCA for decomposing the spectral data since it allows for the inclusion of the y-outlier test and also since it is a logical approach if our intention is to remove spectra that are detrimental to calibration model performance. The performances of the individual tests discussed previously were examined and the results are discussed below. Table 1 summarises the results of the individual outlier tests based on PLS and PCA.

Table 1. Summary of results of individual outlier detection tests based on PLS and PCA. **I** indicates outliers wrongly classified as normal observations and **II** indicates normal observations which have been wrongly classified as outliers.

	PLS		**PCA**	
Test	**Outliers Identified**	**Misclassified Outliers**	**Outliers Identified**	**Misclassified Outliers**
"Known"	1-3, 39-43		1-3, 39-43	
Scores Confidence Limits LV1	2, 3, 40-43	**II:** 1, 39	2, 3, 40-43	**II:** 1, 39
Scores Confidence Limits LV2	3, 40-43, 94, 97	**I:** 94, 97 **II:** 1, 2, 39	42, 43, 59	**I**: 59 **II:** 1-3, 39, 40,41
Classical Mahalanobis	None	**II:** All Known outliers	None	**II:** All known outliers
Modified Mahalanobis	2, 3, 40-43	**II:** 1, 39	2, 3, 40-43, 79, 89, 92	**I:** 79, 89, 92 **II:** 1, 39
Hotelling T^2	2, 3, 40-43	**II:** 1, 39	3, 40-43, 67, 72, 77, 79, 89, 92	**I:** 67, 72, 77, 79, 89, 92 **II:** 1, 2, 39
Q residuals	76, 77, 87-92, 96, 98, 99	**I:** All identified outliers **II:** All known outliers	32, 40, 70, 91, 96, 97, 99	**I:** 32, 70, 91, 96, 97, 99 **II**: 1-3, 39, 41-43
Q and T^2	None	**II:** All	40	**I:** None **II:** 1-3, 39, 40-43
Y outliers	3, 42, 43, 91	**I:** 91 **II:** 1, 2, 39-41	N/A	N/A

5.1 ***Scores Confidence Limit Test***

Figure 3 shows the results applying the scores confidence test to the dataset. The horizontal lines indicate the 95% confidence interval calculated using (1). Observations which fall outside this interval are considered as outliers. From Figure 3(a), it can be seen that the first LV picks up 6 observations as outliers (Observations 2, 3, 40-43). It does not pick up the "known" outliers 1 and 39 but there is no false identification. The second LV picks up 7 observations as outliers (Figure (3(b)) namely 3, 40-43, 94 and 97. Only 5 out of the 8 known outliers are detected and spectra 94 and 97 are misclassified as outliers.

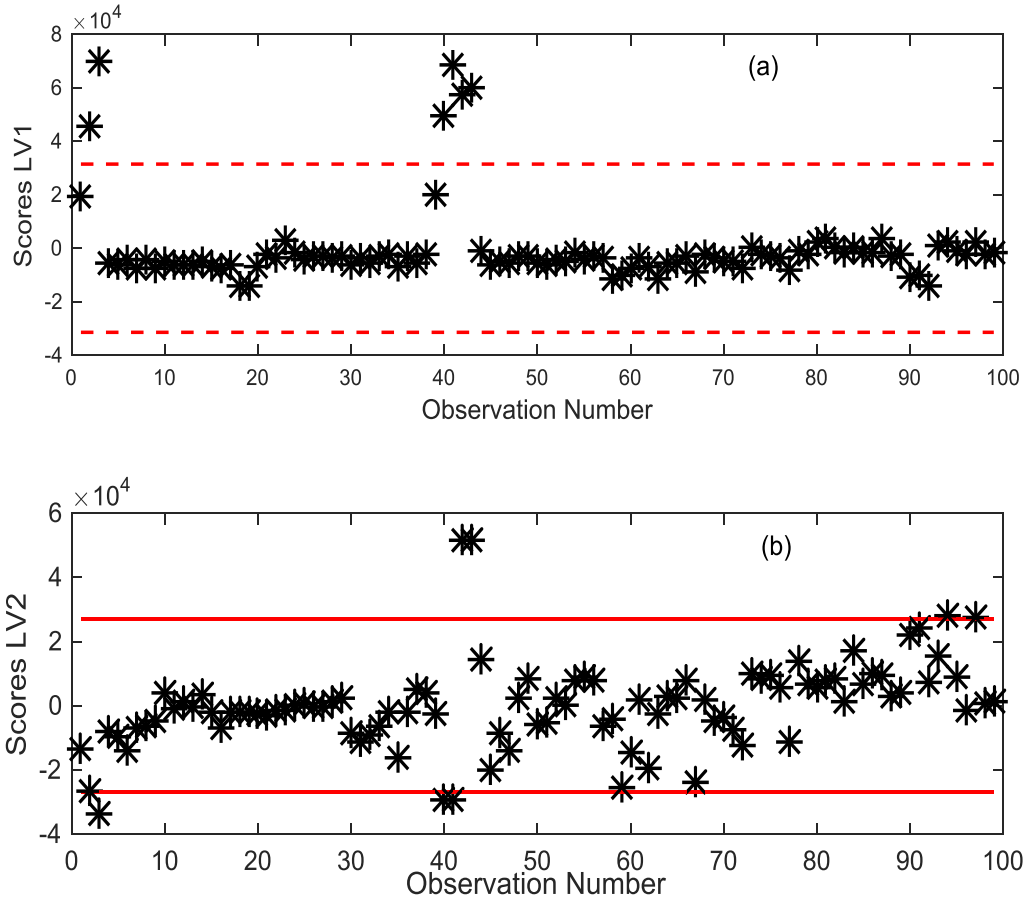

Figure 3. Scores of the first two latent variables for the Raman data with 95% confidence limits (red lines). PLS scores of (a) first latent variable; (b) Second latent variable

When PCA was used instead of PLS to compute the scores (data not shown), the first LV identified the same set of spectra as outliers as when PLS was used. The second LV picked only 3 spectra, 42, 43 and 59 as outliers.

While it is possible to use more LVs in this fashion, the chances of false identification can increase. Thus using 2 or 3 LVs will be a conservative approach to reduce false identification though this could reduce the sensitivity in identifying the true outliers.

5.2 ***Modified Mahalanobis Distance***

Figure 4 shows the results of applying the classical (Figure 4(a)) and modified (Figure 4(b)) Mahalanobis Distance test. Observations that fall outside the 95% confidence limit calculated using (6), given by the horizontal line, are considered as outliers. The conventional MD measure does not identify any outliers with the observations falling well within the 95% confidence limit. It can be seen that the modified MD identifies 6 out of the 8 known outliers namely, 2, 3, 40-43 and does not wrongly classify any of the normal observations.

When PCA was used, the robust MD measure identified the following as outliers: 2, 3, 40-43, 79, 89 and 92 thus identifying 5 of the 8 known outliers and wrongly classifying 79, 89 and 92 as outliers. As was the case with PLS, the classical MD did not identify any of the outliers.

The difference in performance between the classical and modified MD is caused by the use of the variance (in the case of the Classical Mahalanobis Distance) and the interquartile range (in the case of the modified Mahalanobis distance). With the presence of extreme outliers in the data the variance will be much larger than the actual value. Since the variance is in the denominator, this impacts the calculated MD by reducing its value. This has the effect of creating a smaller spread of the values. The interquartile range on the other hand is not

affected by the magnitude of the extreme observations and therefore tends to be smaller in magnitude compared to the variance. This results in a smaller value in the denominator which results in a bigger spread in the MD value. Since the critical Chi-Squared values (see (6)) are the same for both cases, the bigger spread in the MD values translates into higher sensitivity to outlying data.

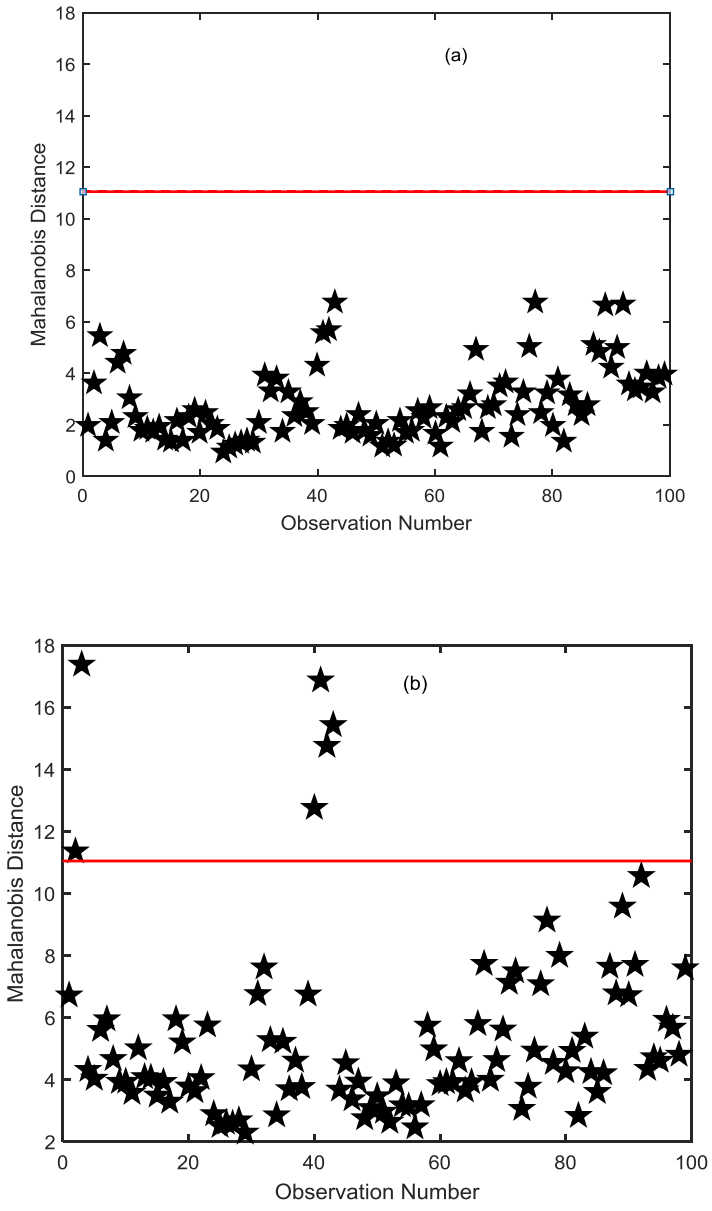

Figure 4. Mahalanobis distances of samples using (a) Classical MD and (b) Modified MD measures.

5.3 ***Leverage Outlier Test***

Figure 5 shows the results for the leverage outlier test. The 95% confidence limits for Q (horizontal line) and T^2 (vertical line) were calculated using (12) and (11) respectively. If the Q residual on its own was used to identify outliers, a total of 11observations will have been classified as outliers when PLS is used. The observations which are classified as outliers are 76, 77, 87-89, 92, 96, 98, 99. None of the 8 known outliers were detected. When PCA was used 7 spectra were identified as outliers: 32, 40, 66, 70, 91, 96, 99. Only one observation, 40, is correctly identified as an outlier. Therefore for this dataset, using the residuals to identify the outliers is not effective.

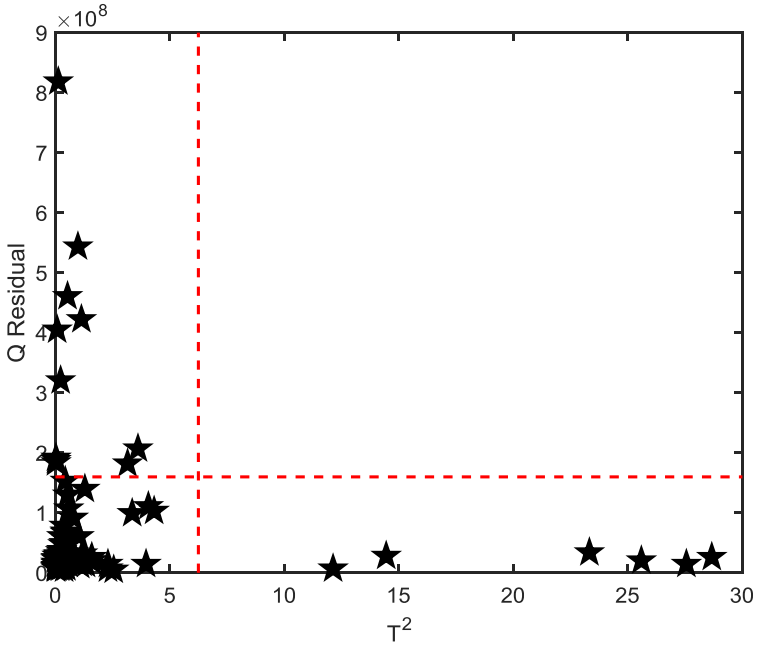

Figure 5. Q residuals vs. Hotelling T^2 values of the observations. The vertical red line is the 95% confidence limit for T^2 and the horizontal line is the 95% confidence limit for the Q residuals.

If T^2 on its own was used to identify outliers, 6 observations would have been identified as outliers if PLS were used: 2, 3, 40-43. Thus, for this dataset, the T^2 test would identify 6 out of the 8 known outliers and there are no falsely classified normal spectra. Using PCA (data not shown), 11 spectra are classified as outliers: 2, 3, 40-43, 79, 89, 92. As in the case of PLS, the

same 6 out of the 8 outliers are identified but PCA leads to 8 normal spectra being misclassified as outliers. Combining the Q and T^2 tests together, since there are no common points identified as outliers, the overall leverage test identifies none of the outliers when PLS is used. PCA leads to 1 outlier, spectrum 40, which is a known outlier.

This investigation of the 3 X-block outlier tests suggests that using PLS could lead to more reliable outlier detection outcomes than PCA and therefore justifies the use of PLS instead of PCA for identifying outliers.

5.4 ***Y Outlier Test***

Figure 6 shows the results for the Y estimate test applied. The horizontal lines indicate the 95% confidence interval calculated using (15). The test identifies 4 observations as being outliers: 3, 42, 43 and 91. Thus 3 out of the 8 known outliers are identified and one normal observation, 91, is misclassified as an outlier.

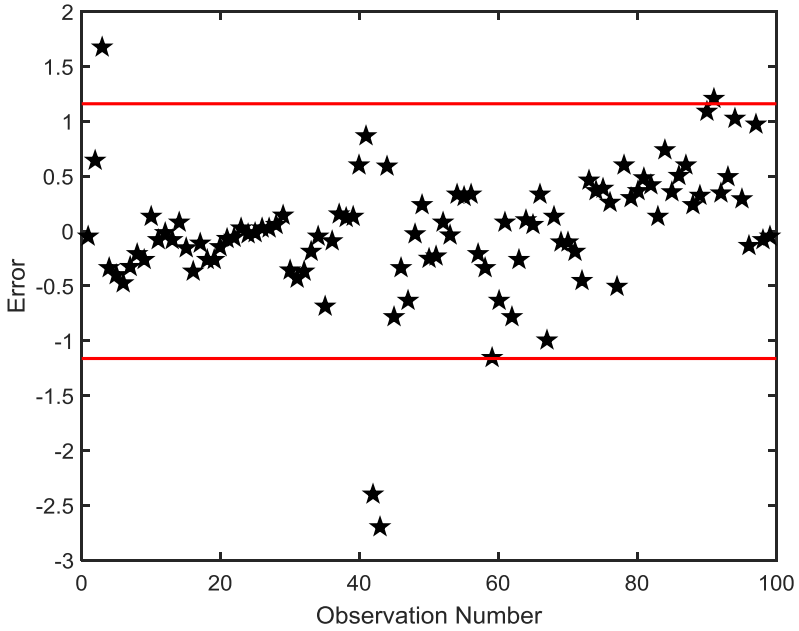

Figure 6. Error in estimation of concentration of Component A. Red lines indicate the 95% confidence bounds which is used for determining whether an observation is an outlier.

The analysis of the individual outlier tests shows that none of them identify all the outliers and some of the tests result in wrongly classifying some normal observations as outliers (See Table 1). We can expect the performance of these individual methods to vary from one dataset to another.

Table2. Output from the automatic outlier detection algorithm is shown for $\gamma = 1$. The trimming step at which the desirability index is at a maximum is shown in bold font. Based on the desirability index, 6 outliers will be removed in total i.e. outliers identified up to iteration 6 are removed.

Trimming Step	Observation Removed	Number of LVs	RMSECV	F	D
0	0	2	0.72	405	0.265
1	3	3	0.67	281	0.186
2	43	2	0.68	266	0.167
3	41	2	0.68	200	0.125
4	42	2	0.18	1592	0.825
5	40	2	0.18	1095	0.557
6	**2**	**4**	**0.07**	**1570**	**0.826**
7	89	4	0.08	1545	0.751
8	1	6	0.05	1649	0.806
9	39	6	0.05	1162	0.597
10	92	7	0.04	1473	0.699
11	48	7	0.03	1724	0.705
12	91	8	0.03	1453	0.582
13	90	9	0.03	1010	0.423
14	49	10	0.03	1127	0.394
15	47	10	0.02	1249	0.340
16	38	10	0.02	1347	0.270
17	95	10	0.02	1449	0.190
18	59	10	0.02	1548	0.100
19	87	10	1.87	1416	0.000

*5.5 **Automated Weighted Outlier Detection***

Table 2 shows the output from the automated outlier method for the dataset considered in this study. In this case, γ was set to 1. The desirability index has a maximum value at trimming step 6 (shown in bold), thus indicating that 6 outliers will be removed by the

automated outlier detection algorithm. These are the observations that are removed at each step up to step 6. The outliers removed are 2, 3, 40-43 which are the same as the ones identified by the modified Mahalanobis measure and the Hotelling T^2 (See Table 1) in a single step. There are no misclassifications. If sequential multivariate trimming is used, we would have reached the maximum trimming level i.e. 19 observations would have been classified as outliers.

Table 3. The effect of the magnitude of the sensitivity factor γ on the number of observations removed as outliers.

γ	**No. of Outliers Removed**	$\mathbf{D_{max}}$
1	6	0.826
0.9	8	0.865
0.8	8	0.923
0.7	11	0.995
0.6	11	1.091
0.5	11	1.189
0.4	11	1.286
0.3	11	1.383
0.2	18	1.539
0.1	18	1.718
0	18	1.90

The impact of the sensitivity parameter on the number of outliers removed can be seen by examining Table 3 in conjunction with Table 1. From Table 3, it can be seen that reducing γ can increase the sensitivity of the algorithm to outliers. However, this can result in some observations being misclassified as outliers. In situations where it is more important to identify the outliers, a lower value for γ is suggested. It can be seen that when the value is between 0.3 – 0.7, the number of outliers identified remains stable at 11. In this case, all the outliers identified by personnel by manual analysis, have been captured by the algorithm. However, it identifies 3 more outliers namely, observations 48, 89 and 92. The results suggest that a wide range of γ can be used by the algorithm to identify approximately the same

outliers that will be chosen by manual analysis. The risk of misclassifying an observation or missing an outlier can be balanced by adjusting γ.

6. CONCLUSIONS

A novel approach to automatically identify outliers has been proposed and demonstrated on Raman spectroscopy data obtained from an industrial distillation process. The results from this comprehensive approach suggest that this method can provide similar outcomes which would be obtained by analysis and decision inputs from data analysts. Apart from the overall methodology, this work introduces several novelties. The system uses PLS instead of PCA which is normally used for detecting multivariate outliers. Analysis indicates that for calibration or recalibration purposes, multivariate outlier detection based on PLS may be more advantageous than PCA. A simple modification to Mahalanobis distance was also proposed which appears to be more sensitive to outliers than the conventional Mahalanobis distance. The methodology also introduces the concept of a desirability function to enable automatic decision making based on multiple statistical measures for outlier detection. A simple desirability function given by (18) for choosing the set of observations as outliers in the calibration dataset was considered. Analysis indicates that the sensitivity parameter γ can be tuned to make the automated outlier detection system more or less aggressive in terms of outlier removal. While this methodology has been considered in this study in the context of calibration and re-calibration, it can be extended for online outlier or change detection applications.

ACKNOWLEDGEMENTS

The authors would like to thank BP for providing PhD studentship funding for Mark Dewar. The authors would also like to thank the Engineering and Physical Sciences Research Council (EPSRC) for providing Doctorial Training Partnership funding.

REFERENCES

[1] Bakeev, K., (Ed.), (2010). Process Analytical Technology. Second Edition. John Wiley and Sons, United Kingdom.
[2] Küppers, S., (2014). Applications of Optical Spectroscopy to Process Environments, in Handbook of Spectroscopy: Second, Enlarged Edition (eds G. Gauglitz and D. S. Moore), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany. doi: 10.1002/9783527654703.ch41.
[3] Hawkins, D., (1980). Identification of Outliers. Chapman and Hall. London.
[4] Ben-Gal I., (2005) "Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, chapter 1, Outlier Detection.
[5] Iglewicz, B. & Hoaglin, D. (1993), "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.
[6] Lu C., Chen D., Kou Y., (2003) "Algorithms for spatial outlier detection," In Proceedings of the 3rd IEEE International Conference on Data-mining (ICDM'03), Melbourne, FL.
[7] Shekhar S., Lu C. T., Zhang P., "Detecting Graph-Based Spatial Outlier," Intelligent Data Analysis: An International Journal, 6(5), 451–468.
[8] Wilson, Paul W (1993). "Detecting outliers in deterministic nonparametric frontier models with multiple outputs." Journal of Business & Economic Statistics 11.3: 319-323.
[9] Penny, K. I. & Jolliffe, I. T. (2001). A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data. Journal of the Royal Statistical Society. 50 (3), 295–307.
[10] Davies L., Gather U., (1993) "The identification of multiple outliers," Journal of the American Statistical Association, 88(423), 782-792.
[11] Acuna E., Rodriguez C. A., (2004) "Meta-analysis study of outlier detection methods in classification," Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Venice.
[12] Hodge, V. J. & AUSTIN, J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review. 22 (2), 85–126.
[13] Adams, M. J (2004). Chemometrics in Analytical Spectroscopy. 2nd ed. Cambridge: The Royal Society of Chemistry. 2-9.
[14] Mahalanobis, Prasanta Chandra (1936). "On the generalised distance in statistics". Proceedings of the National Institute of Sciences of India 2 (1): 49–55
[15] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, (2000) The Mahalanobis distance, Chemometrics and Intelligent Laboratory Systems, Volume 50, Issue 1, Pages 1-18.
[16] Penny, K. I. and Jolliffe, I. T. (1999). Multivariate Outlier Detection Applied to Multiply Imputed Laboratory Data. Statistics in medicine. 18, 1879-1895.
[17] K.V. Mardia, J.T. Kent, J.M. Bibby. "Multivariate Analysis", (1979) Academic Press, Chap 2. P39.
[18] N.K. Shah, P.J. Gemperline Combination of the Mahalanobis distance and residual variance pattern recognition techniques for classification of near-infrared reflectance spectra. Anal. Chem., 62 (1990), pp. 465–470.
[19] H. Liu, Q. Weng. Enhancing temporal resolution of satellite imagery for public health studies: A case study of West Nile Virus outbreak in Los Angeles in 2007. Remote Sens. Environ., 117 (2012), pp. 57–71.
[20] Y. Lu, P.-Y. Liu, P. Xiao, H.-W. Deng Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays Bioinformatics, 21 (2005), pp. 3105–3113.
[21] B. Reiser. Confidence intervals for the Mahalanobis distance. Comm. Statist. Simulation Comput., 30 (2001), pp. 37–45.

[22] J. Hardin and D. M. Rocke.(2005) "The Distribution of Robust Distances", Journal of Computational and Graphical Statistics, Volume 14, Number 4, Pages 1–19.

[23] Dimitrios Ververidis and Constantine Kotropoulos. Gaussian Mixture Modeling by Exploiting the Mahalanobis Distance. IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 56, NO. 7, JULY 2008. 2797-2811.

[24] Rousseeuw, P. J. and Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association. 85 (411), 633-639.

[25] Rousseeuw, P. J. (1985), "Multivariate Estimators With High Breakdown Point," in Mathematical Statistics and its Applications (vol. B), eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht : Reidel, pp. 283–297.

[26] Daniel Peña and Francisco J. Prieto. Multivariate Outlier Detection and Robust Covariance Matrix Estimation. TECHNOMETRICS, AUGUST 2001, VOL. 43, NO. 3 pp 286-310.

[27] K. Ro, C. Zou, Z. Wang. (2015). "Outlier detection for high-dimensional data", Biometrika, 102, 3, pp 589-99.

[28] Upton, G. & Cook, I. (1996). Understanding Statistics. Oxford University Press. p.55.

[29] Hotelling, H., (1931) "The generalization of Student's ratio," The Annals of Mathematical Statistics, vol. 2, no. 3, pp. 360–378.

[30] Wise, B. M., Gallagher, N. B., Bro, R., Shaver, J. M., Windig, W. & Koch, R. S (2006). Chemometrics Tutorial for PLS_Toolbox and Solo. 3905 West Eaglerock Drive, Wenatchee, WA 98801 USA: Eigenvector Research, Inc. 102-159.

[31] Sullivan, J. H. and Woodall, W. H. (1996) "A comparison of multivariate control charts for individual observations," Journal of Quality Technology, vol. 28, no. 4, pp. 398–408.

[32] Vargas, A. N., (2003) "Robust estimation in multivariate control charts for individual observations," Journal of Quality Technology, vol. 35, 4, pp. 367–376.

[33] Jensen, W. A., Birch, J. B. and Woodall, W. H. (2007) "High breakdown estimation methods for phase I multivariate control charts," Quality and Reliability Engineering International, vol. 23, 5, pp. 615–629.

[34] Shabbak, A. and Midi, H.. (2012). An Improvement of the Hotelling T^2 Statistic in Monitoring Multivariate Quality Characteristics. Mathematical Problems in Engineering. Volume 2012 (531864),15.

[35] J. F. MacGregor, T. Kourti. STATISTICAL PROCESS CONTROL OF MULTIVARIATE PROCESSES. Control Engineering Practice. 3 (1995) 403–414.

[36] J. Edward Jackson and Govind S. Mudholkar. Technometrics, Vol. 21, No. 3 (Aug., 1979), pp. 341-349.

[37] Rousseeuw, Peter J."Robustness and outlier detection in chemometrics." Critical reviews in analytical chemistry 36.3-4 (2006): 221-242.

[38] Del Castillo, E. and Montgomery, D. C., (1993), "A Nonlinear Programming Solution to the Dual Response Problem," Journal of Quality Technology, Vol. 25, No. 3.

[39] Derringer, G., and Suich, R., (1980), "Simultaneous Optimization of Several Response Variables," Journal of Quality Technology, 12, 4, 214-219.

[40] Chakravorty, R., Gauri, S.K., and Chakraborty, S. (2013). Optimization of multiple responses of ultrasonic machine (USM) process: A comparative study. International Journal of Industrial Engineering Computations, 4, 285 – 296.

[41] Costa, N. R., Lourenço, J. and Pereira, Z. L. (2011). Desirability function approach: A review and performance evaluation in adverse conditions. Chemometrics and Intelligent Laboratory Systems. 107, 234-244.

[42] Ch'ng, S. and Quah, H. L (2005), A new approach for multiple-response optimization Qual. Eng., 17, 621–626.

[43] Svante Wold. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics, Vol. 20, No. 4, Part 1 (Nov., 1978), pp. 397-405.

[44] Baibing Li , Julian Morris, Elaine B. Martin Chemometrics and Intelligent Laboratory Systems 64(1):79-89 · October 2002.

[45] B. Walczak 1 D.L. Massart. Robust principal components regression as a detection tool for outliers. Chemometrics and Intelligent Laboratory Systems 27 (1995) 41-54.

[46] Egan, W. J. & Morgan, S. L. (1998). Outlier Detection in Multivariate Analytical Chemical Data. Analytical Chemistry. 70 (11), 2372-2379.

[47] N. R. Draper, H. Smith (1966). Applied Regression Analysis. John Wiley and Sons. Chap 1. P 25.