Evaluation of Diagnosis Methods in PCA-based Multivariate Statistical Process Control

Marta Fuentes-García; Gabriel Maciá-Fernández; José Camacho

Dpt. of Signal Theory, Telematics and Communications, CITIC-UGR, University of Granada, 18071, Granada, Spain

Abstract

Multivariate Statistical Process Control (MSPC) based on Principal Component Analysis (PCA) is a well-known methodology in chemometrics that is aimed at testing whether an industrial process is under Normal Operation Conditions (NOC). As a part of the methodology, once an anomalous behaviour is detected, the root causes need to be diagnosed to troubleshoot the problem and/or avoid it in the future. While there have been a number of developments in diagnosis in the past decades, no sound method for comparing existing approaches has been proposed. In this paper, we propose such a procedure and use it to compare several diagnosis methods using randomly simulated data and from realistic data sources. This is a general comparative approach that takes into account factors that have not previously been considered in the literature. The results show that univariate diagnosis is more reliable than its multivariate counterpart.

Key words: MSPC, Diagnosis, Contribution Plots, PCA, Networkmetrics, Smearing

1 Introduction

Multivariate Statistical Process Control (MSPC) based on multivariate methods such as Principal Component Analysis (PCA) is a well-known methodology in the chemometrics domain. This methodology is based on decomposing the data collected from a process into a model and the residual sub-space. Decomposed data are then used to compute a pair of statistics, namely the D-statistic for the model space and the Q-statistic for the residual space. The D-statistic and the Q-statistic are monitored in a pair of Shewhart charts [1] where certain control limits are defined. Anomalous events are detected when these statistics exceed the control limits for a given number of consecutive sampling times in either of the charts [2][3].

Once an anomaly is detected, the related variables should be identified. This process is termed the diagnosis of the fault, and it helps the analysts to identify the root cause of the anomaly so that problems within the process are timely identified and can be corrected. This is why diagnosis is an essential part of MSPC [4]. Despite this relevance and the existence of alternative techniques, no investigation has compared the performance of these methods.

There are several papers in the literature related to diagnosis methods in MSPC. These works introduce new contributions and provide limited comparisons with various reference methods. Some examples are *Contribution Plots* (*CP*) [4][5][6], *Reconstruction-Based Contributions* (*RBC*) [7][8], *Diagonal Contributions* (*DC*) [8] and *Relative Contributions* (*RC*) [8]. Several variations of *Reconstruction-Based Contributions* can be found in the litera-

Email address: `nmfuentes@ugr.es, tel: +34 958 241717` (Marta Fuentes-García).

ture [9][10][11]. All the preceding methods are based on computing the contributions of the variables to the statistics used in PCA-MSPC. An alternative approach is pursued by the *observation-based Missing-data method for Exploratory Data Analysis (oMEDA)* [12], which computes the contributions of the variables to the scores in a given sub-space instead of the contributions to the statistics.

Alcala and Qin [8] present an analysis of Contribution Plots, Reconstruction-Based Contributions and Diagonal Contributions methods. These authors provide a generalization on Contribution Plots, named Complete Decomposition Contributions (CDC) and Partial Decomposition Contribution (PDC). The general expression is called General Decomposition Contribution (GDC).

In [7], the authors perform a Monte Carlo simulation to build a process model to validate the RBC proposal. The experiment is performed using only one data structure given by that simulation. The same structure is also used in [8], and [13][14]. The authors in works [13][14] analyse the smearing effect on Contribution Plots based methods. Similar approaches are followed in [9][10][11]. This reduced structure, and the apparent lack of consideration of the variety of parameters that could affect the diagnosis, such as the number of variables and the variation in the selected PCs, compromises the generality of the results.

An alternative approach to traditional diagnosis methods was presented in [15]. This is based on the study of the correlation between variables. By analyzing changes on the correlations they are able to identify the variables more frequently involved in a fault. The work includes two cases of study based on an artificial network and on the simulation of a Continuous Stirred-Tank Reactor (CSTR) system. The authors study a wide range of data sets and perform a Montecarlo simulation to estimate calibration curves.

Other data driven approaches suggest the application of causal maps to identify the root causes of an anomalous observation. The authors present an evaluation of the methods proposed in their works, including several cases of study on a Tennessee Eastman plant simulation [16][17].

In general, multivariate diagnosis methods suffer from the *smearing* problem: misdiagnosis owing to the spread of the contribution from the variables affected by an anomaly to those not affected by it [14][18]. The *smearing* problem is mainly a result of the correlation between variables that is taken into account when the contributions are computed from the statistics, and the original values of the variables need to be retrieved from only a few values, corresponding to the calculated scores. This turns the considered system into a non-determined system. When a problem occurs and alters the normal value of a variable, even while keeping the remaining variables under control, it causes contamination in other contributions. After detecting the first high contributions in the residual, the model is no longer valid and the scores and residuals cannot be verified [18]. This problem results in a more complex diagnosis process and reinforces the necessity of a comprehensive study comparing these techniques.

In this paper, we present a methodology that enables consideration of the different factors that might influence the performance of diagnosis methods, providing a comparison framework. The methodology has been validated considering different parameters that are varied across a number of synthetic simulations and realistic experiments. Three multivariate diagnosis methods have been included: *CP*, *RBC*, and *oMEDA*. Additionally, we propose a direct univariate diagnosis method derived from the *oMEDA* expression. We term this method *Univariate-Squared* (*U-Squared*).

The rest of the paper is organized as follows. In Section 2, the notation used in this paper is presented. In Section 3, the bases of MSPC, PCA, anomaly detection and diagnosis methods are introduced, and the *Univariate-Squared* method is presented as an alternative for diagnosis. Section 4 introduces a new methodology for the experimental comparison of diagnosis methods. In Section 5, this methodology is applied to several simulated and real data sets and the results are discussed. Finally, in Section 6, the main ideas and contributions of this investigation are summarized.

2 Notation

In this paper, these notation criteria are followed: scalars are specified with lowercase letters, vectors with bold lowercase letters and matrices with bold uppercase letters. If there is no explicit specification, vectors are row vectors by default. Transposed matrices are indicated with an apostrophe. Constants are specified with uppercase letters. Equations presenting matrix and vectorial products and sums of scalars are used indistinctly throughout the paper for clarity.

3 Diagnosis in PCA-MSPC

3.1 PCA-MSPC

MSPC is a methodology to distinguish common from special causes of variation in a process. Essentially, this means discriminating between events that are considered normal in the process and those that seldom occur, *i.e.*, those that are due to an abnormal event. PCA-MSPC is performed using a pair of complementary statistics that enable the indirect monitoring of a high number of variables. The statistics are computed from the PCA decomposition of calibration data to build a model of normal operation [4][5][6]. This methodology is applied to detect whether the behaviour of the incoming data from a process fit the previously calibrated model.

MSPC is applied in two steps:

- *phase I)* It consists of detecting, diagnosing and correcting for special causes of variation in the process, so that only common causes of variation remain. In many cases, e.g. [6], phase I is limited to the removal of outliers, under the belief that the rest of collected data represent an stable process.
- *phase II)* It is performed on new incoming data from the process to detect excursions from the NOCs in a timely manner. When an anomaly is detected, diagnosis is performed to identify its causes and classify its nature [6][19][20].

3.2 PCA-based model

Given an $N \times M$ data matrix, with N the number of observations and M the number of variables, PCA identifies the sub-space with maximum variance in the M-dimensional variables space. With PCA, the original variables are linearly transformed into principal components (PCs). From the PCs, the first A components are selected, capturing most percentage of the variance. The PCA model can be expressed as follows [4][5]:

$$\mathbf{X} = \mathbf{T}_A \cdot \mathbf{P}'_A + \mathbf{E} \tag{1}$$

with $\mathbf{T}_A$ the *score* matrix of size $N \times A$, $\mathbf{P}_A$ the *loading* matrix of size $M \times A$, and $\mathbf{E}$, the *residual* matrix of size $N \times M$, corresponding to the variance not captured by the PCA model.

The scores for a new observation, $\mathbf{t}_{new}$, are computed by projecting the row vector corresponding to that observation, $\mathbf{x}_{new}$, onto the model subspace:

$$\mathbf{t}_{new} = \mathbf{x}_{new} \cdot \mathbf{P}_A \tag{2}$$

Once the scores have been computed, the residuals, $\mathbf{e}_{new}$, are calculated:

$$\mathbf{e}_{new} = \mathbf{x}_{new} - \mathbf{t}_{new} \cdot \mathbf{P}'_A \tag{3}$$

3.3 Anomaly detection

Both scores and residuals are monitored in the MSPC system using two statistics, namely, the D-statistic (D), and the Q-statistic, (Q).

The D-statistic is computed to monitor the model subspace [4][5][18].

$$D_{new} = \sum_{a=1}^{A} \left(\frac{t_a^{new}}{s_a}\right)^2 = \sum_{a=1}^{A} \frac{(t_a^{new})^2}{\lambda_a} \tag{4}$$

where t_a^{new} and s_a^2 are, respectively, the *score* for the a^{th} component and the sample variance of this score. The variances of the principal components are the eigenvalues, λ_a, of $\mathbf{\Lambda} = \frac{1}{N-1} \cdot \mathbf{T}'_A \cdot \mathbf{T}_A$. Where Λ is a diagonal matrix with the A first eigenvalues of $\frac{1}{N-1} \cdot \mathbf{X}' \cdot \mathbf{X}$.

To monitor the residual, the Q-statistic is calculated as:

$$Q_{new} = \sum_{m=1}^{M} (e_m^{new})^2 \tag{5}$$

where e_m^{new} is the residual value of the new observation corresponding to the m^{th} variable.

If any of the statistics corresponding to a new observation is greater than a threshold, termed *Upper Control Limit* (*UCL*), this observation is identified as anomalous. The scores are linear combinations of the original variables; therefore, according to the Central Limit Theorem, they are supposed to follow an approximately Normal distribution [6]. As a consequence, the D-statistic times a constant in phase I follows a beta distribution [3]:

$$D \sim \frac{(N-1)^2}{N} B_{A/2,(N-A-1)/2} \tag{6}$$

Therefore, the corresponding Upper Control Limit (UCL) for the D-statistic at significance level α is given by:

$$UCL(D)_\alpha = \frac{(N-1)^2}{N} B_{(A/2,(N-A-1)/2),\alpha} \tag{7}$$

For new incoming data in phase II, the D-statistic times a constant follows an F distribution [3]:

$$D \sim \frac{A \cdot (N^2-1)}{N \cdot (N-A)} F_{A,(N-A)} \tag{8}$$

And the corresponding UCL at significance level α is given by:

$$UCL(D)_\alpha = \frac{A \cdot (N^2 - 1)}{N \cdot (N - A)} F_{(A,(N-A)),\alpha} \tag{9}$$

Several procedures can be used to determine the UCL for the Q-statistic. Again, the residuals can be assumed to follow a multivariate normal distribution. Jackson and Mudholkar showed in [2] that an approximate critical value at significance level α is given by:

$$UCL(Q)_\alpha = \theta_1 \cdot \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \tag{10}$$

where $\theta_n = \sum_{a=A+1}^{rank(\mathbf{X})} (\lambda_a)^n$, with $rank(\mathbf{X})$ the rank of the matrix of data $\mathbf{X}$ and λ_a the eigenvalues of matrix $\frac{1}{N-1} \cdot \mathbf{E}' \cdot \mathbf{E}$, with $\mathbf{E}$ the matrix of residuals; $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$; and z_α is the $100 \cdot (1 - \alpha)\%$ standardized normal percentile.

Alternatively, the approximation based on the weighted chi-squared distribution proposed by Box can be used [21]. Control limits for the Q-statistic that distinguish phase I and phase II can also be found in [19].

To achieve adequate performance of the monitoring charts in phase II, it is highly recommended to readjust the control limits using the calibration data on a leave-one-out basis [22][23]. The limits are raised or lowered so that the Overall Type I (OTI) risk equals the imposed significance level α. Following the definition in [6], the OTI is the percentage of false alarms in the NOC calibration observations:

$$OTI = 100 \cdot \frac{nf}{N} \% \tag{11}$$

where nf is the number of false alarms (*i.e.*, single observations where the statistic computed surpasses the control limit) in the NOC calibration data.

3.4 Diagnosis methods

In this paper, three existing multivariate methods are selected for comparison. Additionally, a fourth method, corresponding to a univariate approach, is included:

i) Contribution Plots (CP). This is currently the most accepted approach for diagnosis in PCA-MSPC [4][5][6][18].

The contribution of the m^{th} variable to the D-statistic, c_m^D, is obtained from the following expression:

$$c_m^D = \mathbf{t}_{new} \cdot \mathbf{\Lambda}^{-1} \cdot \mathbf{p}'_m \cdot x_m^{new} \tag{12}$$

where $\mathbf{p}_m$ is the vector in the m^{th} row of the *loading* matrix for the A selected PCs, and $\mathbf{\Lambda} = \frac{1}{n-1} \cdot \mathbf{T}'_A \cdot \mathbf{T}_A$ is a diagonal matrix with the A first eigenvalues of $\frac{1}{n-1} \cdot \mathbf{X}' \cdot \mathbf{X}$.

The contribution of the m^{th} variable to the Q-statistic, c_m^Q, corresponding to the residual, is calculated applying

$$c_m^Q = (x_m^{new} - \mathbf{p}_m \cdot \mathbf{t}'_{new})^2 \tag{13}$$

ii) Reconstruction-Based Contributions (RBC). This is a popular method that follows an alternative approach to compute the contributions of the variables to a given statistic [7][8]. It is based on the work of Dunia *et al.* [24].

The contribution rbc_m^D to the D-statistic, corresponding to the model, is calculated from the expression

$$rbc_m^D = \frac{(\mathbf{i}_m \cdot \mathbf{D}_A \cdot \mathbf{x}'_{new})^2}{d_{mm}} \tag{14}$$

where $\mathbf{D}_A = \mathbf{P}_A \cdot \mathbf{\Lambda}^{-1} \cdot \mathbf{P}'_A$, $\mathbf{i}_m$ stands for the m^{th} row vector of the identity matrix, $\mathbf{I}$, with size $M \times M$, and d_{mm} is the m^{th} element in the main diagonal of matrix $\mathbf{D}_A$.

The contribution rbc_m^Q to the Q-statistic, corresponding to the residual, is obtained from

$$rbc_m^Q = \frac{(\mathbf{i}_m \cdot \mathbf{C}_R \cdot \mathbf{x}'_{new})^2}{c_{mm}^R} \tag{15}$$

with $\mathbf{C}_R = \mathbf{P}_R \cdot \mathbf{P}'_R$, where $\mathbf{P}_R$ is the *loading* matrix with the residual components from $A + 1$ to the rank of the data, and c_{mm}^R is the m^{th} element in the main diagonal of matrix $\mathbf{C}_R$.

iii) observation-based Missing-data method for Exploratory Data Analysis (oMEDA). This is a method that was originally designed for exploratory data analysis to compute the contribution of a variable to specific artefacts, such as clusters or outliers, in the scores distribution [12]. Unlike the previous methods, it uses the same expression for the model and residual sub-spaces and it does not compute the contributions to the statistics. Moreover, it simultaneously considers all the observations in the data matrix.

Let us consider the column vector, $\mathbf{x}_m$, containing each observation in the data matrix for the m^{th} variable:

$$\mathbf{x}_m = \hat{\mathbf{x}}_m(Z) + \mathbf{e}_m(Z) \tag{16}$$

where $\hat{\mathbf{x}}_m(Z)$ is the estimated value of $\mathbf{x}_m$ in a given sub-space Z and $\mathbf{e}_m(Z)$ is the corresponding residual. Then, oMEDA follows:

$$d_m^Z = 2 \cdot \mathbf{x}'_m \cdot \mathbf{D} \cdot |\hat{\mathbf{x}}_m(Z)| - \hat{\mathbf{x}}'_m \cdot \mathbf{D} \cdot |\hat{\mathbf{x}}_m(Z)| \tag{17}$$

that can be re-expressed only in terms of $\mathbf{x}_m$ obtaining the estimated value, $\hat{\mathbf{x}}_m(Z)$, from Equation (16) as follows:

$$d_m^Z = (\mathbf{x}'_m + \mathbf{e}'_m(Z)) \cdot \mathbf{D} \cdot |\mathbf{x}_m - \mathbf{e}_m(Z)| \tag{18}$$

where $\mathbf{D} = \frac{\mathbf{d} \cdot \mathbf{d}'}{\|\mathbf{d}\|^2}$ and $\mathbf{d}$ is a dummy column vector with non-zero values in positions corresponding to the observations to be studied [1] . In this paper, we are interested in the diagnosis of one observation, which is possible by using $\mathbf{d} = 1$, which in turn means, $\mathbf{D} = 1$. Then, the expressions corresponding to the two sub-spaces under study can be obtained by substituying $\mathbf{D} = 1$ in Equation (18):

$$d_m^{\mathcal{D}} = (x_m^{new} + e_m^{new}) \cdot |x_m^{new} - e_m^{new}| \tag{19}$$

where $Z = \mathcal{D}$ refers to the model sub-space, *i. e.*, the corresponding to the selected PCs; and

$$d_m^{\mathcal{Q}} = (x_m^{new} + \hat{x}_m^{new}) \cdot |x_m^{new} - \hat{x}_m^{new}| \tag{20}$$

where $Z = \mathcal{Q}$, refers to the residual sub-space, *i.e.*, the corresponding to the non-selected PCs.

Note that the superscripts $\mathcal{D}$ and $\mathcal{Q}$ are used to maintain the consistency with the terminology used in the previously studied diagnosis methods.

[1] The way of selecting the possible non-zero values is out of the scope of this work, for further details we recommend to follow the original paper [12].

iv) Univariate-Squared (*U-Squared*). The correlation structure in the model may not hold for a detected anomaly and therefore the division in model/residuals found for calibration data may not be optimum for diagnosis. If this occurs, one can consider that it makes no sense to calculate the contribution of the variables to each statistic separately. Under this hypothesis, it might be interesting to take into account the full variable space for diagnosis. The fact that oMEDA is equally computed for the model and residual subspaces makes its extension to the complete variable space possible. Thus, setting $Z = \mathcal{D} + \mathcal{Q}$ and using Equations (19) and (20) we obtain:

$$u_m = d_m^{\mathcal{D}+\mathcal{Q}} = x_m^{new} \cdot |x_m^{new}| \tag{21}$$

which is equivalent to $u_m = sign(\mathbf{x}_m^{new}) \cdot (\mathbf{x}_m^{new})^2$.

Note that this expression, which we have called *Univariate-Squared*, corresponds to a univariate approach because it considers only the original value of each variable and not the scores. Similar approaches have been analysed elsewhere [13][14] but their proficiency has not been proven through a general comparison.

The univariate proposal contrasts with the accepted trend in PCA-based MSPC diagnosis: it adopts a univariate approach although a multivariate detection has been previously applied. This method does not suffer from the smearing problem, as the correlation, which is the main cause of the smearing [13] [14][18], does not affect to the computation. To assess *U-Squared* and to check whether the univariate approach is valid for diagnosis in MSPC, we have included this method in the comparative.

4 A methodology for the comparison of diagnosis methods

A comprehensive comparison methodology should meet several requirements:

- *Generation of anomalies with known diagnosis.* It must be known what variables are affected by the anomaly prior to the comparison to check whether the diagnosis methods correctly identify these variables.
- *Definition of a metric to evaluate the diagnosis performance.* Having a measure that indicates how much the known anomalous variables stand out from the non-affected variables enables assessment of the diagnosis ability of each evaluated method.
- *Evaluation of factors affecting the diagnosis.* The parameters that might have an impact on the methods under consideration must be identified and should be assessed when a comparison is conducted.
- *Extraction of low uncertainty results.* Varying the affecting previously identified parameters and repeating the configuration for different observations or models is one way to obtain sufficiently large experiments, minimizing the uncertainty.

We have developed an MSPC-based methodology to compare diagnosis methods based on these requirements.

4.1 Generating anomalies with known diagnosis

In this paper, we propose artificially generating anomalies by modifying NOC observations to ensure that there are no other anomalies in the observation except those introduced in this way. Let us consider an observation from the

NOC matrix, $\mathbf{x} \in \mathbf{X}_{NOC}$, *i.e.*, an anomaly-free observation. We follow $\mathbf{x}_{alt} = \mathbf{x} + \mathbf{r}$ to obtain an anomalous observation, where

$$x_{alt,m} = \begin{cases} x_m, & \text{if } x_m \in \dot{\mathbf{x}}, \\ \chi \cdot s, & \text{if } x_m \in \tilde{\mathbf{x}}, \end{cases} \tag{22}$$

and $s = sign(x_m)$, and χ is the altered value of the observation for the previously selected set of variables, $\tilde{\mathbf{V}}$, that makes either of the statistics, D or Q, exceed its Upper Control Limit (UCL). $\tilde{\mathbf{x}}$ are the variable(s) to be altered, and $\dot{\mathbf{x}}$ are those that do not modify the original value.

We alter the original value of $\mathbf{x}$ by following Equation (22) until either of the statistics is equal to the corresponding UCL multiplied by a given factor, K, obtaining the anomalous observation, $\mathbf{x}_{alt}$. This can be done in several ways, such as:

- *Trial and error*. We iteratively increase χ to modify the normal value of the selected variables.
- *Analytically.* We use analytic expressions to compute a new value, χ, to alter the selected variables.

The numeric approach can be computationally intensive, and for this reason, we use the analytic expressions provided in Equations (25) and (26).

To derive the analytical expression, let us start by analyzing the equation applied to compute the D-statistic, $Dst = \mathbf{t} \cdot \mathbf{\Lambda}^{-1} \cdot \mathbf{t}'$, where $\mathbf{t}$ is the score for the observation, $\mathbf{x}$, to be altered.

Considering that $\mathbf{t} = \mathbf{x} \cdot \mathbf{P}_A$, the vector can be reordered into affected and non-affected variables: $\mathbf{x}_{alt} = [\dot{\mathbf{x}} \; \tilde{\mathbf{x}}]$ and their corresponding *loading* matrix:

$\mathbf{P}_A = \begin{bmatrix} \dot{\mathbf{P}}_A \\ \\ \tilde{\mathbf{P}}_A \end{bmatrix}$, where $\tilde{\mathbf{x}}$ are the variable(s) to be altered, $\dot{\mathbf{x}}$ are those maintaining their original values, and $\tilde{\mathbf{P}}_A$ and $\dot{\mathbf{P}}_A$ are their loadings. A re-defined expression for the D-statistic is:

$$Dst = (\dot{\mathbf{x}} \cdot \dot{\mathbf{P}}_A + \tilde{\mathbf{x}} \cdot \tilde{\mathbf{P}}_A) \cdot \mathbf{\Lambda}^{-1} \cdot (\dot{\mathbf{P}}'_A \cdot \dot{\mathbf{x}}' + \tilde{\mathbf{P}}'_A \cdot \tilde{\mathbf{x}}') \tag{23}$$

For fixed $\dot{\mathbf{x}}$, solving the equation given by $Dst \;=\; K \;\cdot\; UCL_D$ provides the value for $\tilde{\mathbf{x}}$ from the quadratic expression:

$$Dst = d^D + \tilde{\mathbf{x}} \cdot b^D + \tilde{\mathbf{x}}^2 \cdot a^D = K \;\cdot\; UCL_D \tag{24}$$

with $d^D = \dot{\mathbf{x}}\cdot\dot{\mathbf{P}}_A\cdot\mathbf{\Lambda}^{-1}\cdot\dot{\mathbf{P}}'_A\cdot\dot{\mathbf{x}}'$, $b^D = \mathbf{s}\cdot(2\cdot\tilde{\mathbf{P}}_A\cdot\mathbf{\Lambda}^{-1}\cdot\dot{\mathbf{P}}'_A\cdot\dot{\mathbf{x}}')$, $a^D = \mathbf{s}\cdot\tilde{\mathbf{P}}_A\cdot\mathbf{\Lambda}^{-1}\cdot\tilde{\mathbf{P}}'_A\cdot\mathbf{s}'$, and $\mathbf{s}$ is a vector containing the original sign of each variable in $\mathbf{x}$. Finally, the value for observation $\mathbf{x}$ after applying the alteration, $\mathbf{x}_{alt}$, is obtained by replacing the variables to be altered with the result of solving Equation (24) as a normal quadratic equation and selecting the solution that keeps the sign in the discriminant:

$$\tilde{x}^D = (-b^D + \sqrt{(b^D)^2 - 4 \cdot a^D \cdot c^D})/(2 \cdot a^D) \tag{25}$$

where $\tilde{x}^D$ is the new value assigned to each selected variable for the altered observation, $\mathbf{x}_{alt}$, and $c^D = d^D - K \cdot UCL_D$.

Similarly, to alter a given observation, $\mathbf{x}$, for the Q-statistic it is necessary to replace the equation to solve, $Qst = K \cdot UCL_Q$ and to consider $Qst = \mathbf{t}_R \cdot \mathbf{t}'_R$. This makes $d^Q = \dot{\mathbf{x}} \cdot \dot{\mathbf{P}}_R \cdot \dot{\mathbf{P}}'_R \cdot \dot{\mathbf{x}}'$, $c^Q = d^Q - K \cdot UCL_Q$, $b^Q = \mathbf{s} \cdot (2 \cdot \tilde{\mathbf{P}}_R \cdot \dot{\mathbf{P}}'_R \cdot \dot{\mathbf{x}}')$ and $a^Q = \mathbf{s} \cdot \tilde{\mathbf{P}}_R \cdot \tilde{\mathbf{P}}'_R \cdot \mathbf{s}'$, where $\mathbf{t}_R$ and $\mathbf{P}_R$ stand for the score and the loadings in the residual. The resulting expression is:

$$\tilde{x}^Q = (-b^Q + \sqrt{(b^Q)^2 - 4 \cdot a^Q \cdot c^Q})/(2 \cdot a^Q) \tag{26}$$

Finally, the new value for the variables to be altered is:

$$\chi = \min(\tilde{x}^D, \tilde{x}^Q) \tag{27}$$

4.2 Defining a metric for the diagnosis performance

By defining a ratio based on the relation between the contribution of the anomalous variables and the contribution of the non-affected variables, it is possible to assess and compare the diagnosis power of the methods. We propose a metric calculated as the ratio between the average of the contributions from these variables. We denote this ratio *Diagnosis Goodness Ratio*, γ.

$$\gamma = \frac{\mu_{\tilde{\mathbf{c}}}}{\mu_{\dot{\mathbf{c}}}} \tag{28}$$

with $\mu_{\tilde{\mathbf{c}}}$ and $\mu_{\dot{\mathbf{c}}}$

$$\mu_{\tilde{\mathbf{c}}} = \frac{\sum_{\tilde{x}_m \in \tilde{\mathbf{x}}} |c(\tilde{x}_m)|}{\mathcal{V}} \tag{29}$$

$$\mu_{\dot{\mathbf{c}}} = \frac{\sum_{\dot{x}_m \in \dot{\mathbf{x}}} |c(\dot{x}_m)|}{M - \mathcal{V}} \tag{30}$$

where $c(\tilde{x}_m)$ are the contributions for the modified variables, $c(\dot{x}_m)$ the contributions for the non-affected variables in the altered observation, $\mathbf{x}_{alt} = [\dot{\mathbf{x}}\ \tilde{\mathbf{x}}]$, and $\mathcal{V}$ is the number of altered variables. The greater the ratio γ is, the better the diagnosis power for the method. If the value of γ is close or equal to 1, there is no diagnosis capability.

4.3 Evaluating factors affecting the diagnosis

Our goal is to evaluate how different factors could affect the diagnosis results. Different parameters might be identified as relevant in a comparative study and could then be varied along the scenarios to provide general results. We consider the following factors:

- *Selected PCs* (pcs). How to select the number of PCs to build a PCA-model remains an open problem because the number affects the quality of the model [5][6][25][26].
- *Number of variables to alter* ($\mathcal{V}$). The number of variables affected by an anomaly is important for multivariate diagnosis and it can be varied during the analysis. Note that the expressions proposed in Equations (25) and (26) allow the alteration of any number of variables.
- *Size of the calibration matrix* (τ). This size is the relationship between the *number of variables* (M) to be observed and the *number of samples* (N) to be studied. We consider different types of matrices: *Fat* matrices, F, with $M > N$; *Square* matrices, S, with $N \simeq M$; and *Thin* matrices, T, with $N > M$.

We have selected parameters related both to the detection/diagnosis methods (number of PCs), and also (meta)parameters related to the data under monitoring (type of the calibration matrix and number of variables to be altered).

Note that, to the best of our knowledge, these parameters have not been considered in previous comparisons.

4.4 Yielding low uncertainty results

We follow a Monte Carlo procedure to perform an experimental comparison according to the identified needs, achieving low uncertainty results.

Let us call $\mathbf{X}_{NOC}$ the NOC data set to be altered during the experiment. A *core algorithm*, Algorithm 1, has been designed based on Sections 4.1, 4.2 and 4.3.

The core algorithm is repeated over $\mathcal{Y}$ experiments by considering every combination of the parameters: the type of matrix (τ), the number of selected PCs (pcs), the number of variables to be altered ($\mathcal{V}$), and the randomly generated NOC observations ($nobs$). For each observation $\mathbf{x}$, v random variables are selected to obtain the set of variables $\tilde{\mathbf{V}}$ to be altered, where v varies in the range $\{1 : \mathcal{V}\}$ and corresponds to the number of selected variables. Once the anomaly is generated, it is introduced in these selected variables, producing $\mathbf{x}_{alt}$. Then, the statistics for the anomalous observation are computed, in addition to the number of anomalies on each statistic ($nDst$ and $nQst$), the contributions and the ratios.

5 Experimental section

All the experiments have been implemented using Matlab®. The *MEDA-Toolbox*, a set of multivariate analysis tools for the exploration of data sets [12], has been used to implement the studied methods and the proposed methodology.

Algorithm 1 Comparison of diagnosis methods - Core algorithm

```
procedure core–CMP–Methods
    for each $\tau \in \{T, S, F\}$ do
        for each $pc \in pcs$ do
            for $v \in \{1 : \mathcal{V}\}$ do
                for each observation $n \in \{1 : nobs\}$ do
                    $\mathbf{x} \leftarrow \mathbf{X}_{NOC}(n)$
                    $\mathbf{x}_{alt} \leftarrow \mathbf{x}$ // Anomalous observation is initialized to $\mathbf{x}$
                    $\tilde{\mathbf{V}} \leftarrow \{\tilde{v}_1, ..., \tilde{v}_v\}$ // Select $\tilde{v}_v$ randomly
                    $\tilde{x}^D_{\tilde{\mathbf{V}}} \leftarrow$ Anomaly generation Eq.(25)
                    $\tilde{x}^Q_{\tilde{\mathbf{V}}} \leftarrow$ Anomaly generation Eq.(26)
                    $\chi \leftarrow \min\{\tilde{x}^D_{\tilde{\mathbf{V}}}, \tilde{x}^Q_{\tilde{\mathbf{V}}}\}$ Eq.(27)
                    $\mathbf{x}_{alt}(\tilde{V}) \leftarrow \chi$ // Variables in $\tilde{V}$ take the new value, $\chi$
                    Compute $Dst(\mathbf{x}_{alt})$ and $Qst(\mathbf{x}_{alt})$
                    Increase $nDst$ if $Dst > UCL_D$
                    Increase $nQst$ if $Qst > UCL_Q$
                    for each $method$ do
                        Compute contributions
                        $\gamma \leftarrow \frac{\mu_{\tilde{\mathbf{c}}}}{\mu_{\dot{\mathbf{c}}}}$ // Ratio calculation
                    end for
                end for
            end for
        end for
    end for
end procedure
```

To assess the performance of the selected methods, we have computed and compared the corresponding ratios under a wide range of simulated situations using *simuleMV* [27]. *simuleMV* is simulation software that generates random data for a given level of correlation, $\delta \in \{0, 9\}$, where 0 is applied when there is no correlation and 9 is applied when the correlation is the maximum. The software takes into account the number of observations, N, and the number of variables, M, for the matrix to be simulated. *simuleMV* also enables the generation of a data matrix based on a given covariance matrix.

The results have been validated using two data sets related to real fields of activity: one obtained by simulating the Saccharomyces process (*chemometrics*) [6][28] and the other using traffic data from a communications network (*networkmetrics*) [29].

Note that the Monte Carlo approach allows the generation of anomalies that cover a wide range of possibilities, both univariate and multivariate and both holding/breaking the correlation structure in the model. Unlike in other related works [16][17][30][15] we skip the use of first principles models in the anomaly generation procedure to avoid drawing conclusions that only hold in very specific cases/processes. However, the results should be interpreted considering that there is no theoretical warranty that all types of failure are covered.

We have auto-scaled the data in all cases, as we assume that they come from heterogeneous sources.

τ	N	M	δ	$\mathcal{Y}$	pcs	$\mathcal{V}$	$nobs$
Thin (T)	100	10	$\{3, 6, 9\}$	10	$\{1, 2\}$	$\{1, 2, 3\}$	100
Square (S)	100	100	$\{3, 6, 9\}$	10	$\{1, 4\}$	$\{1, 2, 3\}$	100
Fat (F)	100	1000	$\{3, 6, 9\}$	10	$\{1, 11\}$	$\{1, 2, 3\}$	100

Table 1

Parameters involved in the Monte Carlo Simulation - δ, N and M are parameters in *simuleMV*

5.1 Simulation data sets

Table 1 shows the configuration for the experiment using the methodology proposed in this paper.

- Three types of matrices - T (*Thin*) $= 100 \times 10$, S (*Square*) $= 100 \times 100$, F (*Fat*) $= 100 \times 1000$ - are simulated.
- Three different correlation levels, δ, are considered for each type of matrix: *low* $= 3$, *normal* $= 6$ and *high* $= 9$.
- $\mathcal{Y} = 10$ different models are generated for each type of matrix and correlation level.
- The number of selected PCs is: *i)* $pcs = 1$, and *ii)* the number of PCs that captures the 75% of the total variance.
- The number of variables to be altered, $\mathcal{V}$, is varied from 1 to 3.
- The number of random observations selected is $nobs = N$.

We distinguish between statistics Q and D in the results because this difference has traditionally been considered. According to the expressions defined in Equations (25) and (26), the variables are altered until any of the statistics is

$K = 2$ times its upper control limit. The *core algorithm* is applied iteratively over the presented parameters.

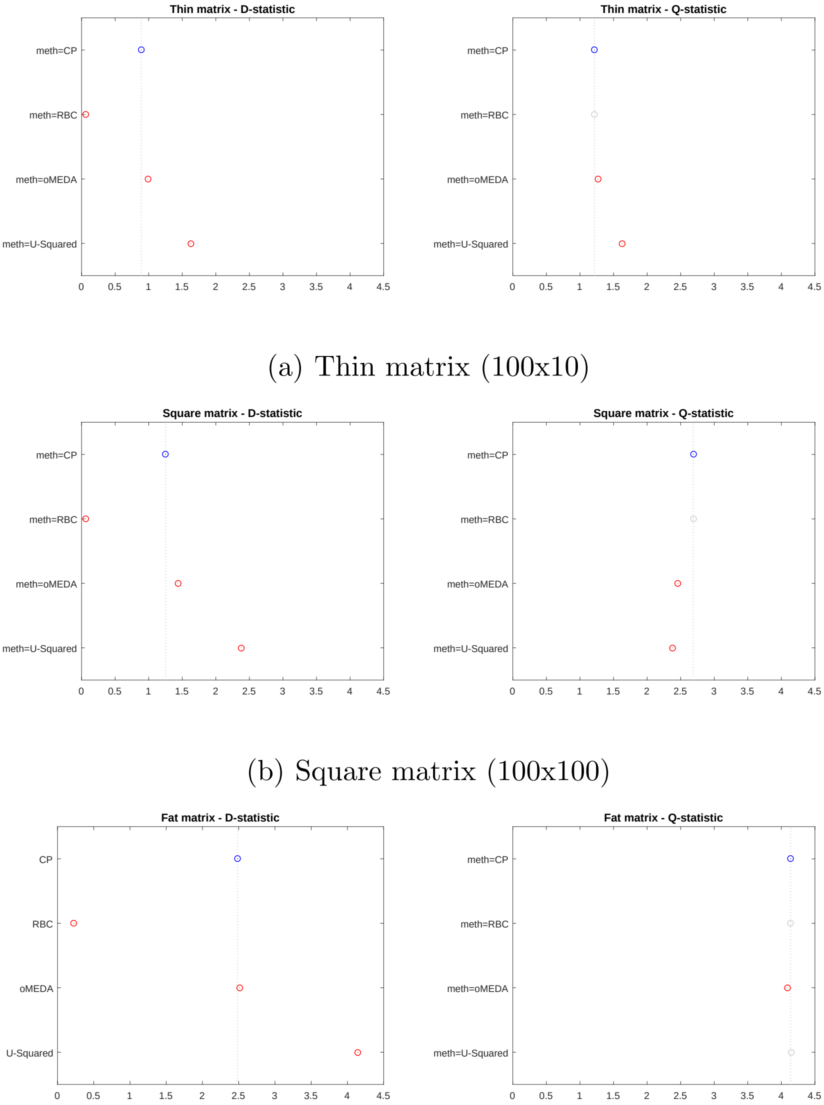
5.1.1 Results

The comparison study includes an analysis of variance (ANOVA) performed on the ratio values. A logarithmic transform is applied to the ratio outcomes to smooth their positive skewness. The test includes the factors of the experiment: correlation level (δ), selected PCs (pcs), number of affected variables ($\mathcal{V}$), diagnosis method, type of matrix (τ), statistics, and first-order interactions. The ANOVA result shows that all these factors and their corresponding interactions, except the correlation level, are statistically significant ($p - value < 0.01$).

We are also interested in identifying which of the studied factors are most relevant. With this aim, the effect size

$$\eta^2 = SS(f)/SS(total) \tag{31}$$

has been computed, where $SS(f)$ is the sum of squares of the evaluated factor, f, according to the ANOVA decomposition, and $SS(total)$ is the total sum of squares. The most relevant parameters, sorted by η^2, are the type of matrix (τ), the statistic, and the diagnosis method. These parameters also present strong interactions; thus, varying any of them has a considerable effect on the other. This suggests that the comparison of the diagnosis methods should be performed individually for each combination of statistic and type of matrix. For that, we compute the least significant difference (LSD) plots, shown in Fig. 1 when statistically significant differences are identified among the approaches. *U-Squared* is in most cases better than the other methods, except



(a) Thin matrix (100x10)

(b) Square matrix (100x100)

(c) Fat matrix (100x1000)

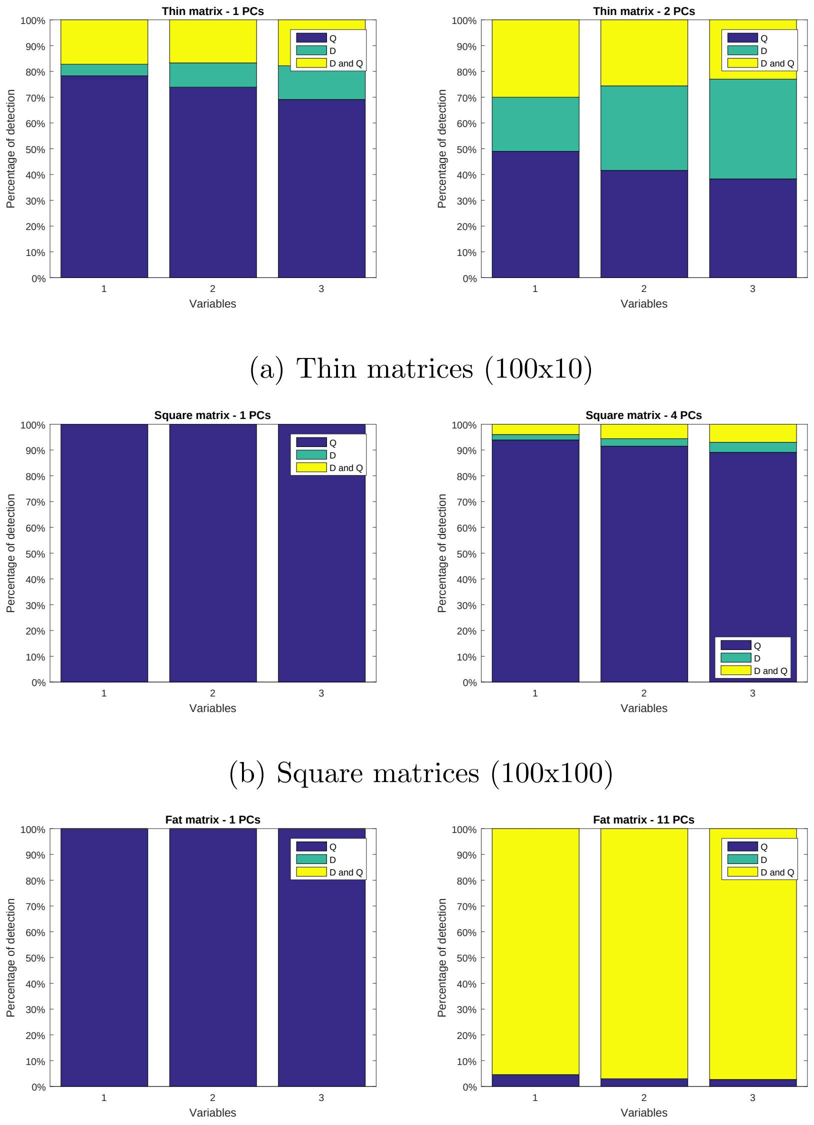
Fig. 1. ANOVA indicates that the results are significant for the selected parameters.

for Square matrices for the Q-statistic, where CP and RBC are better.

As a part of the study of the results, we identify whether faults are detected in the D-statistic, the Q-statistic or both. The percentage of detection for a normal correlation level, $\delta = 6$, is shown in Fig. 2. In general, the probability of detecting an anomaly in the D-statistic increases with the number of affected variables and the number of selected PCs whereas the percentage of detecting an anomaly in the Q-statistic is always greater than that of the D-statistic. Though not shown in the figure, this trend has been observed to be greater or equal when the correlation level is increased. We have also found that in Square matrices with the maximum correlation level, the greatest probability of detecting an anomaly occurs in the D-statistic, and it is very low in the Q-statistic.

To study the distribution in the diagnosis results and to interpret the ANOVA from a practical viewpoint, different plots have been produced using the Advance Blox Plot, *aboxplot* [31]. These plots include the mean value represented by a circle, together with the quartiles and outliers. We have represented the ratios, γ, for each type of matrix classified according to the statistic to which the methods are applied. We have also differentiated the number of selected PCs. Fig. 3 and 4 show the results for a normal correlation level, $\delta = 6$. The outcomes for anomalies detected in the D-statistic are on the left, while those for the Q-statistic are on the right. Note that if an anomaly has been detected in both Q and D, it is in both images. According to the observed percentages of anomaly detection for each statistic, there are only a few or no anomalies in the D-statistic when 1 PC is selected for the Square and Fat matrices. Therefore, those with a detection percentage less than 5% are not considered and only the ratios for the Q-statistic for Square and Fat matrices are shown in Fig. 3 when 1 PC is selected.

The *Univariate-Squared* method is confirmed, in general, to be better or equivalent to the other methods. From a pragmatic view, the difference is relevant for the D-statistic and, for Thin matrices, also for the Q-statistic. Although the differences in the Q-statistic are not important, applying *U-Squared* avoids the smearing effect as it does not take into account the correlation between



(a) Thin matrices (100x10)

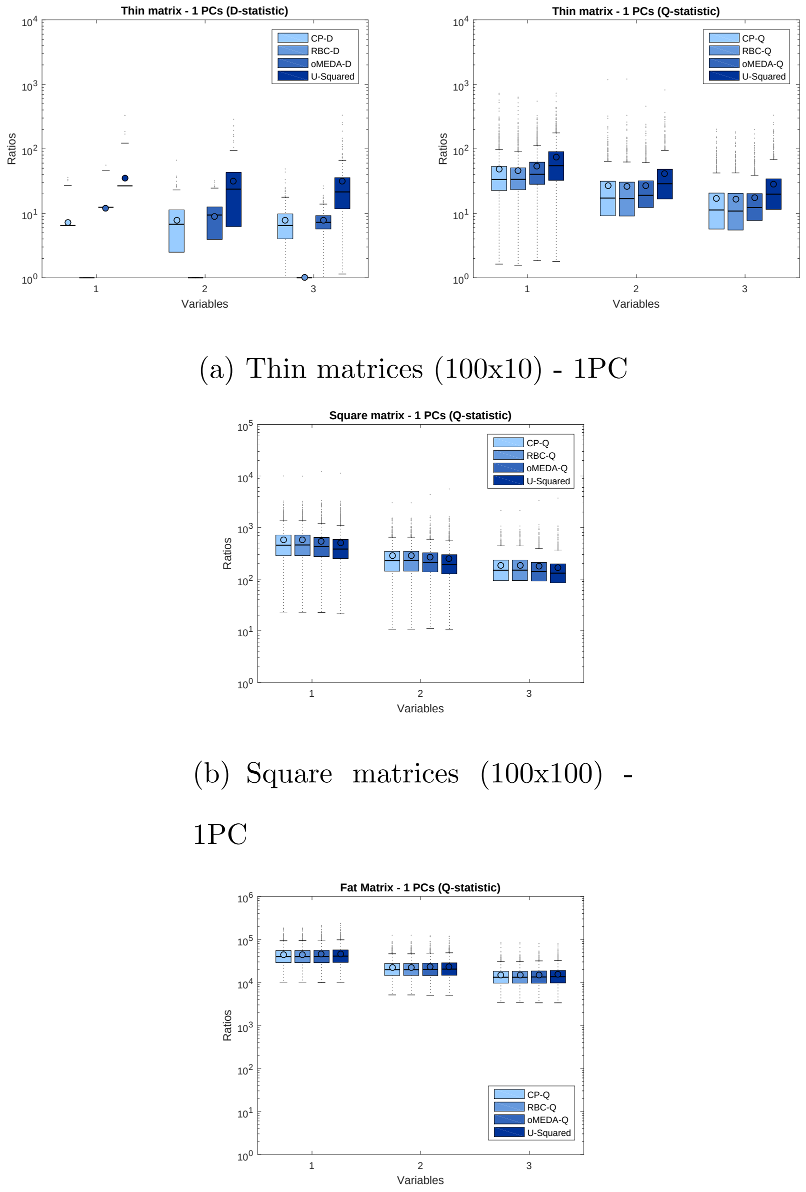
(b) Square matrices (100x100)

(c) Fat matrices (100x1000)

Fig. 2. Percentage of detected anomalies by the statistics for 1 PC and for the number of PCs that captures 75% of the total variance: 2 PCs, 4 PCs, and 11 PCs for (a) Thin matrix (T), (b) Square matrix (S) and (c) Fat matrix (F) simulated with correlation level $\delta = 6$ (normal correlation)

variables. This makes the differences between the results more evident when the number of PCs is increased, which is also the reason for the larger differences in the D-statistic.

The *Reconstruction-Based Contributions* method has a very low ratio when



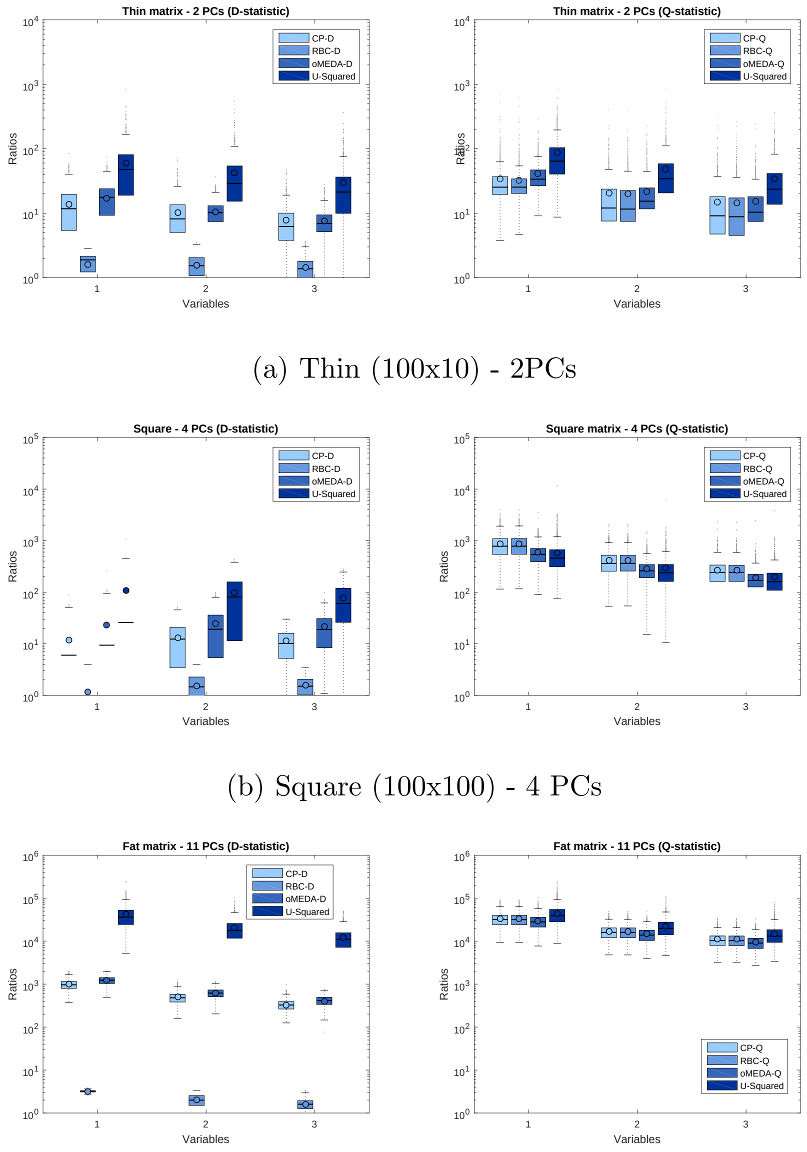
(a) Thin matrices (100x10) - 1PC

(b) Square matrices (100x100) - 1PC

(c) Fat matrices (100x1000) - 1 PC

Fig. 3. Ratios, γ, for 1 PC for (a) Thin matrix (T), (b) Square matrix (S) and (c) Fat matrix (F) simulated with correlation level $\delta = 6$, corresponding to normal or normal interdependence.

the diagnosis is performed for the D-statistic. In fact, when 1 PC is selected, the ratio γ is always equal to 1, indicating a lack of diagnosis capability. This result is mathematically proven in the Appendix of this paper by deriving the RBC expression for the D-statistic.



(a) Thin (100x10) - 2PCs

(b) Square (100x100) - 4 PCs

(c) Fat (100x1000) - 11 PCs

Fig. 4. Ratios, γ, for 2 PCs, 4 PCs, and 11 PCs for (a) Thin matrix (T), (b) Square matrix (S) and (c) Fat matrix (F) simulated with correlation level $\delta = 6$, corresponding to normal or normal interdependence.

Finally, we have verified these results using mean-centered data. We have found that although the ratios are generally lower than auto-scaling, the performance of the methods is the same as when using auto-scaled data.

τ	N	M	$\mathcal{Y}$	pcs	$\mathcal{V}$	$nobs$
Thin (T)	3000	11	1	$\{1,2\}$	$\{1,2,3\}$	3000
Square (S)	900	781	1	$\{1,2\}$	$\{1,2,3\}$	900
Fat (F)	30	1100	1	$\{1,2\}$	$\{1,2,3\}$	30

Table 2

Parameters involved in verification using *Saccharomyces cerevisiae* process data

5.2 Validating the results using realistic data sets

After performing the comparison with simulation data, we have tested whether the results are consistent for data obtained from real applications. With this aim, two additional data sources are considered: one data set obtained by simulating the *Saccharomyces* process (*chemometrics*) [28] and the other corresponding to traffic data from a communications network (*networkmetrics*) [29]. These data sources are considered because *chemometrics* is where PCA-MSPC is most commonly applied [6][32][33] and *networkmetrics* is a growing application area that uses a variation of MSPC, termed MSNM (Multivariate Statistical Process Monitoring) [20] [34].

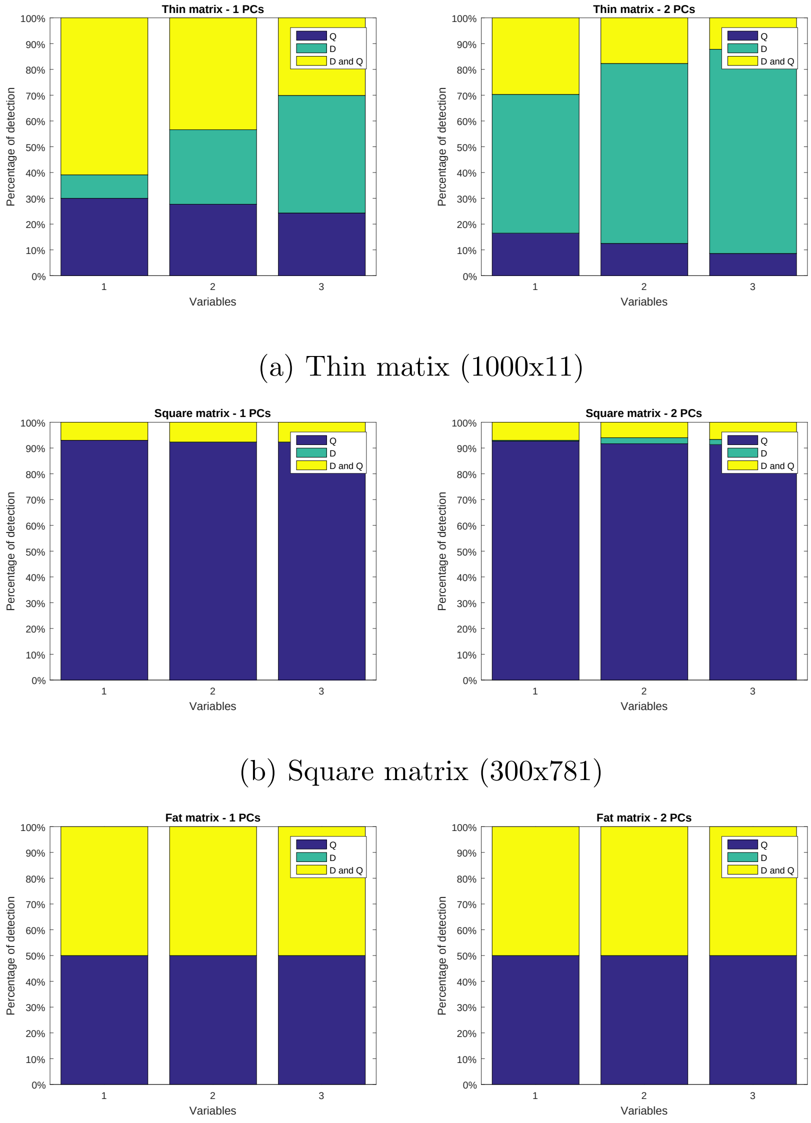
5.2.1 Saccharomyces

This test is based on the *Saccharomyces cerevisiae* batch process [22][28]. As the data are three-way, they have been unfolded for the application of PCA-MSPC. *Batch-wise*, *Variable-wise* and *Batch-Dynamic* unfolding [35] have been used to obtain Fat, Thin and Square matrices, respectively. The parameters for the Monte Carlo experiment are shown in Table 2.

These data have been altered in the same way as in the simulation section but using only one replicate for each considered type of calibration matrix (Thin, Square and Fat). The number of selected PCs is *i)* $pcs = 1$, and *ii)* $pcs = 2$, *i. e.*, the number of PCs that captures the 75% of the total variance. The number of observations selected is $nobs = N$. The variables are altered until either of the statistics is $K = 2$ times its control limit.

The comparative study includes, similarly to *simuleMV*, an ANOVA performed on the ratio values. The test considers the factors of the experiment: selected PCs (pcs), number of affected variables ($\mathcal{V}$), diagnosis method, type of matrix (τ), and statistics, as well as the first-order interactions. The ANOVA result is consistent with that obtained using *simuleMV* as it shows that all these factors and the corresponding interactions are statistically significant ($p - value < 0.01$).

The effect size is also computed using Equation (31) to verify whether the factors identified as relevant in the simulation are valid for the data set from the *Saccharmyces* process simulation. The most relevant parameters are the same as those in the *simuleMV* results: the type of matrix (τ), the statistic, and the diagnosis method. These parameters also present similar strong interactions. According to these results, the comparison is performed individually for each combination of statistic and type of matrix for the diagnosis methods. We compute the least significant difference (LSD) plots when statistically significant differences are identified among the approaches. The results are in agreement with those from the simulation: *U-Squared* is in most cases better than the other methods, except for the Q-statistic for Square matrices, where CP and RBC are better. For the Q-statistic for Fat matrices, there is no difference in any of the methods.



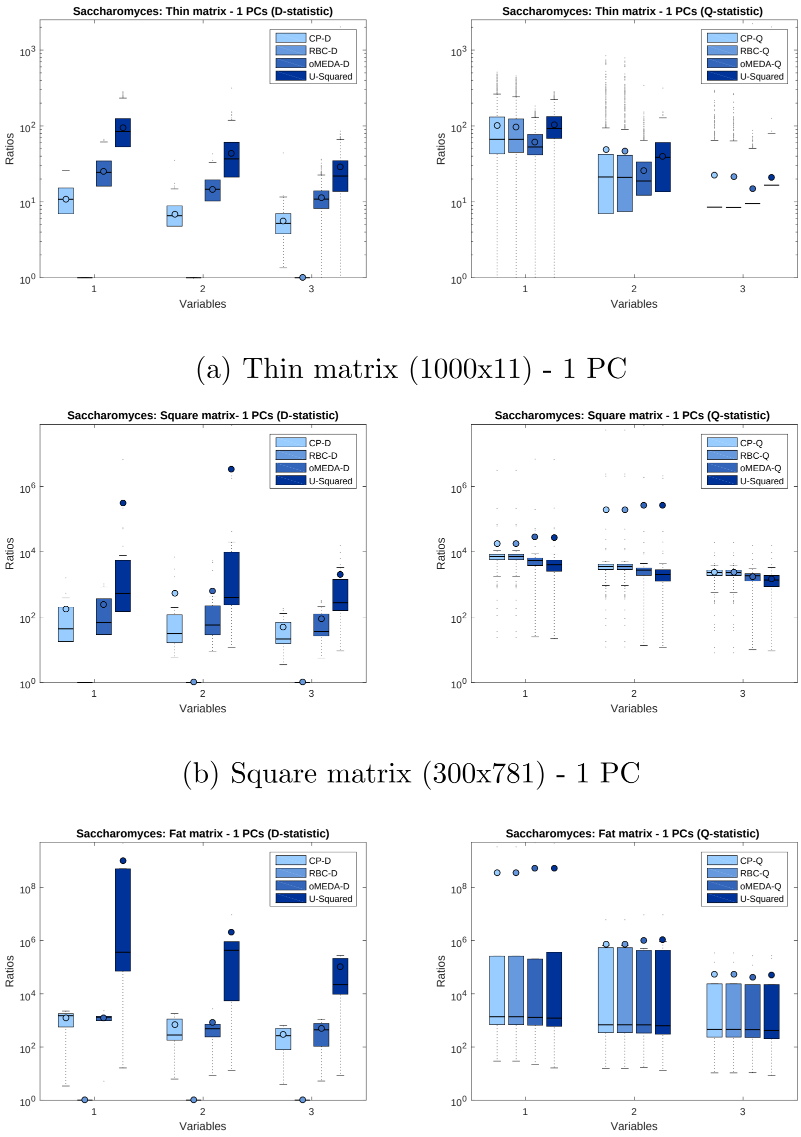
(a) Thin matix (1000x11)

(b) Square matrix (300x781)

(c) Fat matrix (10x1100)

Fig. 5. Percentage anomalies detected by the statistics for 1 PC and 2 PCs for (a) Thin matrix (T), (b) Square matrix (S) and (c) Fat matrix (F) corresponding to the *Saccharomyces cerevisiae* process simulation.

Fig. 5 shows the percentage of anomalies detected for each statistic. Compared to the distribution of probabilities obtained using *simuleMV*, the probability of detection only in the Q-statistic has decreased for each type of matrix. There is a greater probability of detecting an anomaly in both statistics simultaneously, compared to the simulation results. Fig. 6 and Fig. 7 show the



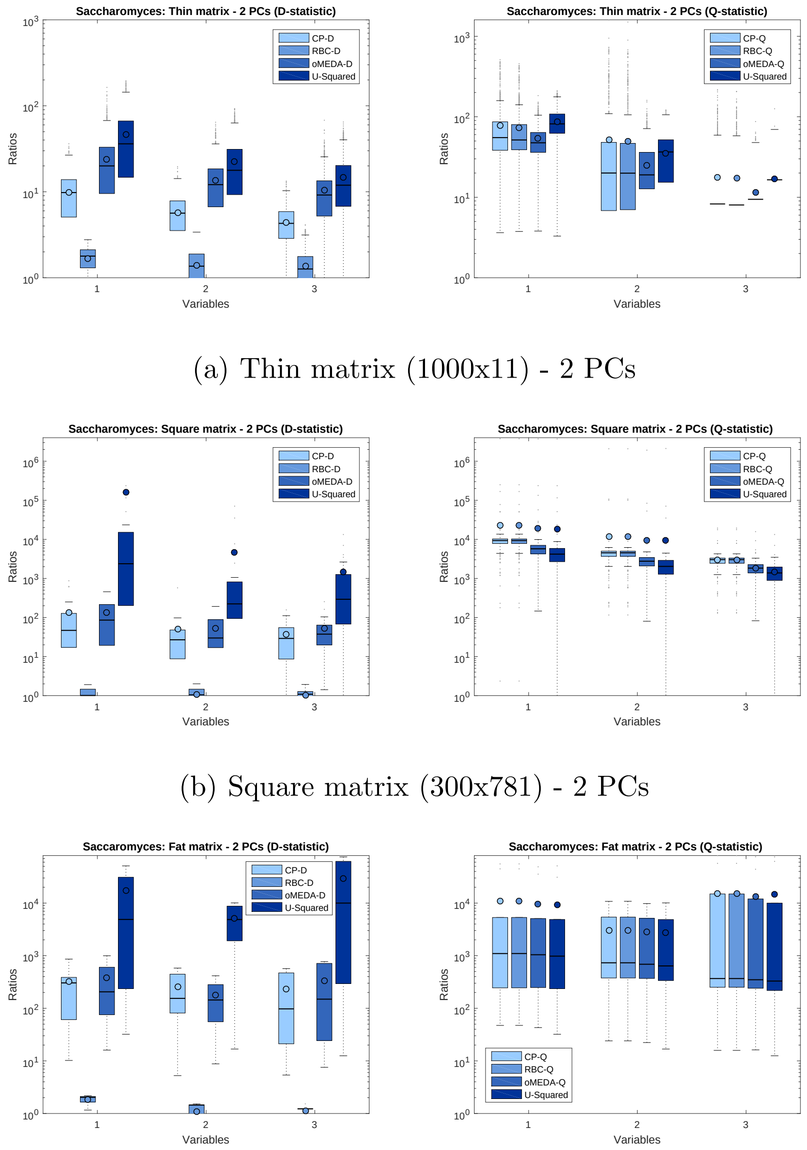
(a) Thin matrix (1000x11) - 1 PC

(b) Square matrix (300x781) - 1 PC

(c) Fat matrix (10x1100) - 1 PC

Fig. 6. Ratios, γ, for 1 PC for (a) Thin matrix (T), (b) Square matrix (S) and (c) Fat matrix (F) corresponding to the *Saccharomyces cerevisiae* process simulation.

distribution of the ratios computed after applying the diagnosis methods. The outcomes for anomalies detected in the D-statistic are on the left, whereas those for the Q-statistic are on the right. Note that if an anomaly is detected in both Q and D, it is included in both graphics. From a practical viewpoint, the differences are more relevant for the D-statistic, and similarly to the simulated data results, these differences are more evident when the number of



(a) Thin matrix (1000x11) - 2 PCs

(b) Square matrix (300x781) - 2 PCs

(c) Fat matrix (10x1100) - 2 PCs

Fig. 7. Ratios, γ, for 2 PCs for (a) Thin matrix (T), (b) Square matrix (S) and (c) Fat matrix (F) corresponding to the *Saccharomyces cerevisiae* process simulation.

PCs is increased. For the Q-statistic, the difference between *U-Squared* and the other methods is not important, although it is a useful way to avoid the smearing effect.

For this data set, RBC does not show good results for the D-statistic and cannot be used for diagnosis when only 1 PC is selected.

τ	N	M	$\mathcal{Y}$	*pcs*	$\mathcal{V}$	*nobs*
Fat (F)	501	24	1	$\{1, 2\}$	$\{1, 2, 3\}$	1000

Table 3

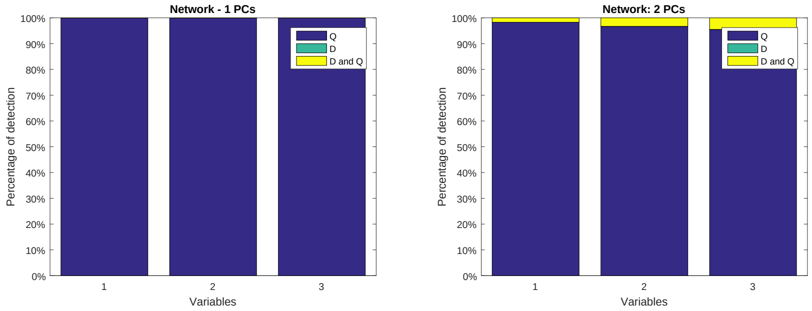
Parameters involved in Verification using *Network data*

5.2.2 Communications Network Traffic

Data from a communications network have been also used in this study. The full data set is the same as used in [29]. This data set has been split into two sets: a calibration set, corresponding to one-third of the observations, **cal**, and a test set with the remaining observations, **X**. The matrix **cal** contains $N = 501$ observations and $M = 24$ variables. **X** includes one hour with network attacks and, in order to avoid polluted values, corresponding observations have been removed. Additionally, only data below 50% of the UCL are used to ensure the test data are NOC. The final data set, $\mathbf{X}_{NOC}$, has $N = 303$ observations and $M = 24$ variables. The type of matrix could be considered *a priori* to be a Thin model, however, its rank is 10, imposed by the rows, which is closer to a Fat or even a Square matrix with dimensions 10 × 24.

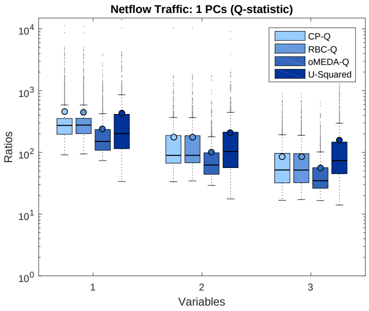
The core algorithm is run using $nobs = 1000$ random observations for only one type of matrix. The number of selected PCs is *i)* $pcs = 1$, and *ii)* $pcs = 2$, *i. e.*, the number of PCs that captures the 75% of the total variance. The variables are altered until either of statistics is $K = 2$ times its control limit. The configuration for the experiment is shown in Table 3.

ANOVA is performed on the ratio values to compare the results with those from the simulated data. The test takes into account the factors of the experiment: selected PCs (pcs), number of affected variables ($\mathcal{V}$), diagnosis method, and statistics, as well as their first-order interactions. Note that the type of



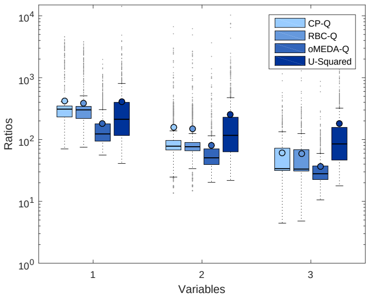
(a) Communications Network Traffic

Fig. 8. Percentage anomalies detected by the statistics for 1 PC and 2 PCs corresponding to *Communications Network Traffic*.



(a) Communications Network Traffic - 1 PC

Fig. 9. Ratios, γ, for 1 PC corresponding to *Communications Network Traffic*.



(a) Communications Network Traffic - 2 PCs

Fig. 10. Ratios, γ, for 2 PCs corresponding to *Communications Network Traffic*.

matrix is not needed in this case. The result from the test is consistent with that obtained using $simuleMV$ and shows that all these factors and the corresponding interactions are statistically significant ($p - value < 0.01$).

The effect size is also computed using Equation (31) to verify whether the factors identified as relevant in the simulation are valid for the considered *communications network* data set. The most relevant parameters are, sorted by η^2, the statistic, the diagnosis method, and the number of altered variables. These parameters are, without considering the type of matrix, the same as those from the simulation and also present strong interactions. Using these results, the comparison is performed individually for the diagnosis methods for each statistic. We compute the LSD plots when statistically significant differences are identified among the approaches. The results are in agreement with those from Fat matrices in the simulation: *U-Squared* is better than the other methods.

Fig. 8 shows the percentages of anomaly detection for each statistic. The probability of detection of an anomaly in only the Q-statistic is closer to that of a Square matrix. Fig. 9 and 10 show the distribution of the ratios computed after applying the diagnosis methods. According to the observed anomaly detection percentages for each statistic, there are only a few or no anomalies in the D-statistic. Therefore, those which have a detection percentage less than 5% are not considered, and only the ratios for the Q-statistic are shown in Fig. 9 and 10. From a practical outlook, there are no significant differences between *U-Squared*, *CP* and *RBC*, but *U-Squared* has the advantage of avoiding the smearing effect.

6 Conclusion

In this paper, we define a methodology to perform experimental comparisons between different diagnosis methods. This methodology satisfies the requirements we previously identified for a comprehensive comparison: *i)* anomalies with known diagnosis are generated from NOC data, *ii)* we define a way to measure the diagnosis ability of each method, *iii)* factors that might affect the diagnosis are identified, and *iv)* we perform a Monte Carlo procedure to obtain low uncertainty results. We believe that this is a generic methodology, since the Monte Carlo approach allows the generation of anomalies that cover a wide range of possibilities, both univariate and multivariate and holding/breaking the correlation structure in the model. However the results should be interpreted considering that there is no theoretical warranty that all types of failure are covered.

Three diagnosis methods of multivariate statistical processes control in the industry are compared using the proposed methodology: Contribution Plots (CP), Reconstruction Based Contributions (RBC) and observation-based Missing-data method for Exploratory Data Analysis (oMEDA). A fourth method that follow a univariate approach is also included, Univariate Squared (*U-Squared*), with the following rationale: when an anomaly breaks the model of normal operation, the decomposition into two subspaces is no longer valid so it makes sense to consider the full variable space for diagnosis. This approach does not suffer from the smearing problem, as the correlation between variables is not considered. Applying oMEDA to the full variable space led us to derive the *U-Squared* expression and to include it in the comparison.

The analysis of variance performed as part of the study indicates that several parameters are relevant to the diagnosis: the dimensions (rows × columns) of the matrix, the diagnosis method, the statistic (D or Q), the number of selected PCs and the number of affected variables. Corresponding least significant difference (LSD) plots show that *U-Squared* is statistically significant better than the other methods. Representing the results with box plots shows these results are especially relevant for the D-statistic. This is because U-squared is computed in the full variable space, avoiding the problems that result from considering the correlation captured for the selected PCs after the occurrence of an anomaly, *i. e.*, the smearing of information among the variables.

The RBC method presents very low ratios for the D-statistic and we mathematically prove, in an Appendix to this paper, that this method does not have diagnosis ability when 1 PC is selected in the D-statistic because all the contributions have exactly the same value.

The comparison is validated using realistic data sets from a *Saccharomyces* process simulation and from a communications network, and the results are consistent with those obtained from the simulated data.

This study has led us to propose a mixed PCA-MSPC process in which detection is performed using a multivariate approach but diagnosis is performed via a univariate method.

7 Acknowledgements

This work is funded by the Ministry of Economy and Competitiveness and FEDER funds through the TIN2014-60346-R project.

Appendix A. Mathematical demonstration of ratio = 1 when 1 PC is selected for the D-statistic and RBC is applied

The RBC expression for the *Hotelling's* T^2 is analysed here. $T^2 = \mathbf{x}' \cdot \mathbf{D}_A \cdot \mathbf{x}$, with $\mathbf{D}_A = \mathbf{P}_A \cdot \boldsymbol{\Lambda}_A^{-1} \cdot \mathbf{P}'_A$, is used to define the RBC expression for the D-statistic in [7][8]. Following a similar derivation procedure as in [36], we can define:

$$\breve{\alpha}_m^A = \sum_{a=1}^{A} \frac{p_{m,a}^2}{\Lambda_a} \tag{32}$$

$$\breve{\beta}_{v,m}^A = \sum_{a=1}^{A} \frac{p_{v,a} \cdot p_{m,a}}{\Lambda_a} \tag{33}$$

where $p_{m,a}$ is the loading of the variable m and the selected component a, Λ_a is the element corresponding to the selected component a on the main diagonal of $\boldsymbol{\Lambda}_A$, $p_{v,m}$ is the loading of variable v, and $\breve{\beta}_{v,m}$ are the elements that do not belong to the main diagonal of the matrix.

Let us consider now Equation (34) to be the expression in the D-statistic for the variable m

$$\mathbf{i}_m \cdot \mathbf{D}_A \cdot \mathbf{x} = \breve{\alpha}_m^A \cdot x_m + \sum_{v \neq m} \breve{\beta}_{v,m}^A \cdot x_v \tag{34}$$

and

$$d_{m,m} = \mathbf{i}'_m \cdot \mathbf{D}_A \cdot \mathbf{i}_m = \breve{\alpha}_m^A \tag{35}$$

the element $d_{m,m}$ corresponding to the diagonal of the matrix $\mathbf{D_A}$. From equations (14) and (35):

$$rbc_m^D = \mathbf{x}' \cdot \mathbf{D}_A \cdot \mathbf{i}_m \cdot (\mathbf{i}'_m \cdot \mathbf{D}_A \cdot \mathbf{i}_m)^{-1} \cdot \mathbf{i}'_m \cdot \mathbf{D}_A \cdot \mathbf{x} \tag{36}$$

is the extended form of Equation (14) for RBC. By combining it with Equation (34), it can be re-written as follows:

$$\begin{aligned} rbc_m^D = {} & \frac{(\check{\alpha}_m^A)^2 \cdot x_m^2 + \sum_{v \neq m} (\check{\beta}_{v,m}^A)^2 \cdot x_v^2}{\check{\alpha}_m^A} + \\ & \frac{2 \cdot \check{\alpha}_m^A \cdot x_m \cdot \sum_{v \neq m} \check{\beta}_{v,m}^A \cdot x_v}{\check{\alpha}_m^A} + \\ & \frac{2 \cdot \sum_{v \neq m} \sum_{w \neq v \neq m} \check{\beta}_{v,m}^A \cdot \check{\beta}_{w,m}^A}{\check{\alpha}_m^A} \end{aligned} \tag{37}$$

By applying Equation (37) for 1 selected PC, and replacing $\check{\alpha}_m^A$ and $\check{\beta}_{v,m}^A$ with Equations (32) and (33), the Equation (38) is obtained:

$$\begin{aligned} rbc_m^D = {} & ((\frac{p_{m,1}^2}{\Lambda_1})^2 \cdot x_m^2 + \sum_{v \neq m} (\frac{p_{m,1} \cdot p_{v,1}}{\Lambda_1})^2 \cdot x_v^2 + \\ & 2 \cdot \frac{p_{m,1}^2}{\Lambda_1} \cdot x_m \cdot \sum_{v \neq m} \frac{p_{m,1} \cdot p_{v,1}}{\Lambda_1} \cdot x_v + \\ 2 \cdot \sum_{v \neq m} \sum_{w \neq v \neq m} & \frac{p_{m,1}^2 \cdot p_{v,1} \cdot p_{w,1}}{\Lambda_1} \cdot x_v \cdot x'_w) \cdot \frac{1}{p_{m,1}^2 / \Lambda_1} \end{aligned} \tag{38}$$

By grouping and simplifying Equation (38) in Equations (39) and (40),

$$\begin{aligned} rbc_m^{D_{1PC}} = {} & \frac{1}{\Lambda_1} \cdot p_{m,1}^2 \cdot x_m^2 + \frac{1}{\Lambda_1} \cdot \sum_{v \neq m} p_{v,1}^2 \cdot x_v^2 + \\ & \frac{2}{\Lambda_1} \cdot x_m \cdot \sum_{v \neq m} p_{m,m} \cdot p_{v,1} \cdot x_v + \\ & \frac{2}{\Lambda_1} \cdot \sum_{v \neq m} \sum_{w \neq v \neq m} p_{v,1} \cdot p_{w,1} \cdot x_v \cdot x'_w \end{aligned} \tag{39}$$

$$rbc_m^{D_{1PC}} = \frac{1}{\Lambda_1} \cdot \sum_v p_{v,1}^2 \cdot x_v^2 + \\ \frac{2}{\Lambda_1} \cdot \sum_v \sum_{w \neq v} p_{v,1} \cdot p_{w,1} \cdot x_v \cdot x_w' = \\ rbc_v^{D_{1PC}} \quad (40)$$

it is shown that the RBC value for the expression in the D-statistic is exactly the same for every variable, *i. e.*, each variable has the same contribution, which makes, according to Equation (28), the ratio $\gamma = 1$. This is translated into a lack of diagnosis ability for RBC for the D-statistic if 1 PC is selected, as it cannot be distinguished which variables are affected when there is an anomaly.

References

[1] W. A. Shewhart, Economic Control of Quality of Manufactured Product, Reissued by the American Society for Quality.

[2] J. E. Jackson, G. S. Mudholkar, Control procedures for residuals associated with Principal Component Analysis., Technometrics 21 (1979) 331–349.

[3] N. D. Tracy, J. C. Young, R. L. Mason, Multivariate Control Charts for Individual Observations, Journal of Quality Technology 24 (2) (1992) 88–95.

[4] T. Kourti, J. F. MacGregor, Multivariate SPC methods for process and product monitoring, Journal of Quality Technology 28 (4).

[5] B. M. Wise, N. L. Ricker, D. F. Veltkamp, B. R. Kowalski, Theoretical basis for the use of principal component models for monitoring multivariate processes, Process Control and Quality 1 (1) (1990) 41–51.

[6] P. Nomikos, J. MacGregor, Multivariate Statistical Process Control Charts for Monitoring Batch Processes (1995).

[7] C. F. Alcala, S. J. Qin, Reconstruction-based contribution for process monitoring, Automatica 45 (7) (2009) 1593–1600.

[8] C. F. Alcala, S. Joe Qin, Analysis and generalization of fault diagnosis methods for process monitoring, Journal of Process Control 21 (3) (2011) 322–330.

[9] G. Li, C. F. Alcala, S. J. Qin, D. Zhou, Generalized reconstruction-based contributions for output-relevant fault diagnosis with application to the Tennessee Eastman process, IEEE Transactions on Control Systems Technology 19 (5) (2011) 1114–1127.

[10] Y. C. Pan, Y. Dong, S. J. Qin, Fault Diagnosis Using Concurrent Projection to Latent Structures, IFAC-PapersOnLine 48 (8) (2015) 1276–1281.

[11] G. Li, S. J. Qin, T. Chai, Multi-directional reconstruction based contributions for root-cause diagnosis of dynamic processes, Proceedings of the American Control Conference (2014) 3500–3505.

[12] J. Camacho, Observation-based missing data methods for exploratory data analysis to unveil the connection between observations and variables in latent subspace models, Journal of Chemometrics 25 (11) (2011) 592–600.

[13] P. V. D. Kerkhof, J. Vanlaer, G. Gins, J. F. M. V. Impe, Analysis of smearing-out in contribution plot based fault isolation for Statistical Process Control, Chemical Engineering Science (2013) 285–293.

[14] P. V. D. Kerkhof, J. Vanlaer, G. Gins, J. F. M. V. Impe, Contribution plots for Statistical Process Control : analysis of the smearing-out effect, European Control Conference (ECC) (2013) 1–6.

[15] T. J. Rato, M. S. Reis, On-line process monitoring using local measures of association. Part II: Design issues and fault diagnosis, Chemometrics and Intelligent Laboratory Systems 142 (142) (2015) 265–275.

[16] L. H. Chiang, B. Jiang, X. Zhu, D. Huang, R. D. Braatz, Diagnosis of multiple and unknown faults using the causal map and multivariate statistics., Journal of Process Control 28 (142) (2015) 27–39.

[17] L. H. Chiang, R. D. Braatz, Process monitoring using causal map and multivariate statistics: Fault detection and identification, Chemometrics and Intelligent Laboratory Systems 65 (142) (2003) 159–178.

[18] J. A. Westerhuis, S. P. Gurden, A. K. Smilde, Generalized contribution plots in multivariate statistical process monitoring, Chemometrics and Intelligent Laboratory Systems 51 (2000) 95–114.

[19] A. Ferrer, Latent Structures-Based Multivariate Statistical Process Control: A Paradigm Shift, Quality Engineering 26 (1) (2014) 72–91.

[20] J. Camacho, A. Pérez-Villegas, P. García-Teodoro, G. Maciá-Fernández, PCA-based multivariate statistical network monitoring for anomaly detection, Computers & Security 59 (2016) 118–137.

[21] G. E. P. Box, Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: Effect of Inequality of Variance in One-Way Classification, The Annals of Mathematical Statistics 25 (1954) 290–302.

[22] J. Camacho, J. Picó, Online monitoring of batch processes using multi-phase principal component analysis, Journal of Process Control 16 (10) (2006) 1021–1035.

[23] H. Ramaker, E. N. M. van Sprang, J. A. Westerhuis, S. P. Gurden, A. K. Smilde, F. H. van der Meulen, Performance assessment and improvement of control charts for statistical batch process monitoring., Statistica Neerlandica 60 (3) (2006) 339–360.

[24] R. Dunia, S. Joe Qin, Subspace approach to multidimensional fault identification and reconstruction, AIChE Journal 44 (8) (1998) 1813–1831.

[25] J. Camacho, A. Ferrer, Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects, Journal of Chemometrics 26 (7) (2012) 361–373.

[26] J. Camacho, A. Ferrer, Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical aspects, Chemometrics and Intelligent Laboratory Systems 131 (2014) 37–50.

[27] J. Camacho, On the Generation of Random Multivariate Data, Chemometrics and Intelligent Laboratory Systems 160 (2016) 40–51.

[28] F. Lei, M. Rotboll, S. B. Jorgensen, A biochemically structured model for Saccharomyces cerevisiae, Journal of Biotechnology 88 (3) (2001) 205–221.

[29] G. Maciá-Fernández, J. Camacho, P. García-Teodoro, R. A. Rodríguez-Gómez, Hierarchical PCA-Based Multivariate Statistical Network Monitoring for Anomaly Detection, International Workshop on Information Forensics and Security.

[30] T. J. Rato, M. S. Reis, On-line process monitoring using local measures of association. Part I: Detection performance, Chemometrics and Intelligent Laboratory Systems 142 (142) (2015) 255–264.

[31] A. Bikfalvi, Advanced box plot for matlab, `http://alex.bikfalvi.com/research/advanced_matlab_boxplot`.

[32] P. Nomikos, J. F. MacGregor, Monitoring batch processes using multiway principal component analysis, AIChE Journal 40 (8) (1994) 1361–1375.

[33] T. Kourti, P. Nomikos, J. F. MacGregor, Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS, Journal of Process Control 5 (4) (1995) 277–284.

[34] J. Camacho, R. Magán-Carrión, P. García-Teodoro, J. J. Treinen,

Networkmetrics: Multivariate Big Data Analysis in the Context of the Internet, Computers & Security 59 (2016) 118–137.

[35] J. Camacho, J. Picó, A. Ferrer, Bilinear modelling of batch processes. Part I: Theoretical discussion, Journal of Chemometrics 22 (5) (2008) 299–308.

[36] J. Camacho, Missing-data theory in the context of exploratory data analysis, Chemometrics and Intelligent Laboratory Systems 103 (1) (2010) 8–18.