Chemometrics and Intelligent Laboratory Systems 118 (2012) 120–126

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A new approach to estimate variable contributions to Hotelling's statistic

Marco V. Cedeño Viteri, Leandro P.F. Rodríguez Aguilar, Carlos R. Alvarez Medina [1], Mabel C. Sánchez *

Planta Piloto de Ingeniería Química (Universidad Nacional del Sur — CONICET), Camino La Carrindanga Km 7, 8000, Bahía Blanca, Argentina

ARTICLE INFO

Article history:
Received 22 March 2012
Received in revised form 14 August 2012
Accepted 16 August 2012
Available online 23 August 2012

Keywords:
T^2-statistic
Contribution plots
Statistical process control
Fault identification
Process monitoring

ABSTRACT

Hotelling's statistic, also called T^2-statistic, is widely used in statistical process control as an extension of the univariate student's chart to reliably detect out of control status in multivariate processes. Although it is a very efficient tool for detection purposes, by itself, it offers no assistance about the origin of the declared faulty status. Several different approaches have been proposed to estimate the variable values' effect on the overall statistic's value. Some of these strategies work in the original measurement space, while others interpret the results coming from the analysis in latent variable spaces using for example Principal Component Analysis or Independent Component Analysis. With the same purpose, we present a novel approach, based on finding the nearest in-control neighbor of the observation point, in this work. The distance between both points is used to determine the contribution of each variable to the out of control state. Those variables whose distance measures exceed a certain threshold value are considered as suspicious. The results of the proposed strategy are compared with those obtained using other strategies that work both the original and latent variable spaces for case studies extracted from the literature.

1. Introduction

The aim of multivariate analysis is to provide a set of statistical analysis tools which are well suited to deal with data when they are collected for more than one variable. Its main advantage is in the treatment of process with correlated variables. There often exist relationships among the variables of a data set, which make them mutually dependent, therefore it is necessary to incorporate such interdependence when analyzing those data sets. Taking into account such behavior is not a trivial task, and it usually makes multivariate analysis quite different in approach and complexity than the corresponding univariate analysis, when only one variable at a time is considered.

In this sense, Hotelling [1] introduced an extension of the student statistic which enables hypothesis testing for the multivariate case. The so called Hotelling's statistic (T^2-statistic) is a quadratic form based upon the Mahalanobis distance. The relationships among variables are taken into account by means of the covariance matrix, which is used to weigh the relative distance between a given point and the sample mean. Hotelling's statistic has been extensively used as a testing tool that enables both, comparison among data populations and classification of observations as coming from one of several populations.

The formula for the T^2-statistic is based on a covariance matrix that is non singular and therefore can be inverted. When two or more observed variables are perfectly correlated or near correlated, the calculation of T^2-statistic fails. For identifying collinearities it is recommended to examine the condition indexes [2] of the sample covariance matrix, defined as the square root of the ratio of the maximum eigenvalue to each of the other eigenvalues. A condition index greater than 30 indicates the presence of a severe collinearity. In this case, one of the variables involved in the collinearity is removed. To determine the set of collinear variables, the linear combination of variables provided by the eigenvector associated to the smallest eigenvalue is examined. After ignoring the variables with small coefficients, the linear relationship between those that are producing the collinearity problem is obtained. Another method to removing collinearities is to re-construct the inverse of the correlation matrix by excluding the eigenvectors corresponding to the near zero eigenvalues. Sample restrictions also originate the singularity of the covariance matrix. Even though different techniques for matrix regularization are proposed to tackle this problem [3], it is necessary to evaluate if the computation of the covariance matrix changes the performance of the statistic.

From the monitoring standpoint, determining whether the process can be considered as in-control at a given time t is just one step in the procedure. Whenever one observation is believed to show an abnormal behavior, all the effort must be oriented towards finding what the root cause of the deviation is.

The activities related to isolating the variables that indicate the faulty state conform what is known as the identification stage, which is frequently performed by calculating the variable contributions to the inflated statistic. The main purpose of evaluating those contributions is to compare the relative influence of each measured variable on the final T^2-statistic value. It is considered that the largest contributions help to reveal the faulty state. Different strategies have been

* Corresponding author. Tel.: +54 291 4861700; fax: +54 291 4861600.
E-mail address: msanchez@plapiqui.edu.ar (M.C. Sánchez).
[1] Permanent address: Bayer Technology Services, 8500 West Bay Rd, Baytown, TX 77520, United States.

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2012.08.004

proposed to assign variable-contribution values to the statistic taking into account the multivariate nature of process data.

Mason et al. [4,5] proposed to decompose the T^2-statistic value as a summation of J independent parts (where J is the number of measured variables). The partitioning begins by selecting any one of the J variables, and calculating the first term of the sum squaring a univariate t statistic for that variable. The jth term of the sum ($j=2, \ldots, J$) corresponds to the square of the jth variable adjusted by the estimates of the mean and standard deviation of its conditional probability distribution given the $(j-1)$ variables previously taken into account in the sum. Since there exists no fixed order for variables, $J!$ different but non-independent partitions can be obtained. As a possible solution for this problem, the authors suggested to focus the interest in only two of those terms for each partition: the one corresponding to the unadjusted contribution of a single selected variable and, the term containing the adjusted contribution of this variable after the adjustment of the $(J-1)$ remaining ones. Nevertheless, when the inspection of this reduced set of terms is not enough to come to a clear conclusion, all significant conditional terms should be compared to a critical value, increasing the complexity of the identification of the source fault.

An alternative straightforward method to decompose the T^2-statistic as a unique sum of variable contributions was presented by Alvarez et al. [6]. This method, called Original Space Strategy (OSS), also provides an explanation about the physical meaning of the negative contributions. If the monitoring statistical technique reveals the process behavior is abnormal, those variable contributions to the inflated T^2-statistic that are greater than a given control-limit, calculated using the reference population, are identified as suspicious. This strategy has been applied for monitoring batch processes [7] and to identify faulty sensors [8].

Among the methods that work in latent variable spaces, it was Jackson [9] who first proposed the decomposition of T^2-statistic into a sum of principal components (PCs) and performed the identification in terms of the weight of each variable in the out-of-control component. However, in most of the industrial applications it is very difficult to associate a physical meaning to each PC and, the variables associated with out-of-control signals cannot be determined easily. Therefore, Miller et al. [10] and MacGregor et al. [11] proposed to evaluate the contributions of each process variable to the scores that are outside of their confidence limits, and then Nomikos [12] presented an approach to calculate the contributions of each process variable to T^2-statistic instead of to the scores, when latent variables cannot be associated to a meaningful group of process variables. Westerhuis et al. [13] extended the theory of contribution plots to latent variable models with correlated scores and, introduced control limits for the contributions that help in finding the variables whose behaviors are different with respect to those contained in the reference data set.

In all the aforementioned methods, the contribution of each variable to the T^2-statistic value is estimated considering that the remaining variables are fixed at their measured values. Therefore, there is a sole parametric curve defining all the possible values of T^2-statistic as function only of the analyzed variable, as it was pointed out by Alvarez et al. [6]. In this work, a new approach devoted to identify the set of measurements that reveal the faulty state and does not impose the aforementioned restriction is presented. The methodology consists in finding the Nearest In-Control Neighbour (NICN) of the observation point by solving a minimization problem. The distance between these points is used to evaluate the relative influence of each measured variable on the T^2-statistic value. Those variables whose distance measures exceed a certain threshold value help to isolate the root cause of the fault, and are considered suspicious.

This work is organized as follows. In Section 2, previous methodologies for calculating variable contributions to the T^2-statistic are briefly reviewed. The proposed approach is described in Section 3. Next, a performance comparison among the new strategy and other existing methodologies is presented. A conclusion section ends the article.

2. Existing Approaches to Estimate Variable Contributions to T^2-statistic

Let us consider a chemical process in which J variables are measured and monitored over time, and let $\mathbf{z}$ be a process observation vector containing all measurements for a given time instant t. The value of T^2-statistic for $\mathbf{z}$ is given by:

$$T^2 = (\mathbf{z}-\bar{\mathbf{z}})^T \mathbf{S}^{-1}(\mathbf{z}-\bar{\mathbf{z}}) \tag{1}$$

where $\bar{\mathbf{z}}$ is the estimate of the population mean ($\boldsymbol{\mu}$) and $\mathbf{S}$ is the estimation for the variance–covariance matrix $\boldsymbol{\Sigma}$. If it is possible to assume that $\mathbf{z}$ follows a normal multivariate distribution ($\mathbf{z}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$), then T^2-statistic follows a $[J(I^2-1)/(I^2-IJ)]F_{J,I-J}$ distribution, where $F_{J,I-J}$ represents the F distribution with J and $(I-J)$ degrees of freedom and I is the number of observations of the reference population.

As can be seen in Eq. (1), the T^2-statistic has a squared form with the minimum at $\mathbf{z}=\bar{\mathbf{z}}$. Since matrix $\mathbf{S}$ is positive semidefinite, all the possible values for $\mathbf{z}$ will generate statistic's values that are greater than or equal to zero. Standardized observation vectors $\mathbf{x}$ are obtained, such that $\mathbf{x}\sim N(\mathbf{0},\mathbf{R})$ where $\mathbf{R}$ is the correlation matrix of the reference population. Therefore T^2-statistic can be re-written as follows

$$T^2 = \sum_{i=1}^{J}\sum_{j=1}^{J} a_{i,j}x_i x_j \tag{2}$$

where a_{ij} denotes the elements of the inverse matrix of $\mathbf{R}$

$$\mathbf{R}^{-1} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,J} \\ \vdots & \ddots & \vdots \\ a_{J,1} & \cdots & a_{J,J} \end{pmatrix}. \tag{3}$$

This particular structure has been exploited in order to estimate the influence of each measurement on the final statistic's value.

Mason et al. [4,5] proposed to decompose T^2-statistic value as a summation of J independent parts:

$$\begin{aligned} T^2 &= t_1^2 + T_{2\bullet 1}^2 + T_{3\bullet 1,2}^2 + T_{4\bullet 1,2,3}^2 + \ldots + T_{J\bullet 1,2,\ldots,J-1}^2 \\ &= t_1^2 + \sum_{j=1}^{J-1} T_{j+1\bullet 1,\ldots,j}^2 \end{aligned} \tag{4}$$

where t_1^2 is the student value of the first selected variable and $T_{j+1\bullet 1,\ldots,j}^2$ is the contribution of the $(j+1)$th variable adjusted by using estimates of the mean and standard deviation of its conditional probability distribution given the j previously considered variables. Since there exists no fixed variable order, it is possible to obtain $J!$ different (but non-independent) partitions for T^2-statistic.

Alvarez et al. [6] presented an alternative straightforward method to decompose the T^2-statistic as a unique sum of variable contributions, as follows:

$$T^2 = \sum_{j=1}^{J} a_{j,j}\left(x_j^2 - x_j^* 2\right) = \sum_{j=1}^{J} c_j \tag{5}$$

$$x_j^* = -\frac{\sum_{\substack{j=1 \\ j\neq k}}^{J} a_{i,j}x_i}{a_{j,j}} \tag{6}$$

where $x_j^*/2$ is the x_j value that minimizes T^2-statistic given the remaining $J-1$ variable values. This decomposition of T^2-statistic also allows to understand the meaning of a negative variable-contribution and to estimate a bound for it. The variable contribution will take negative values if $0\le x_j \le x_j^*$. The minimum contribution value is c_k^{min} that is

located at $x_j = x_j^*/2$. If x_j is out of $0 \le x_j \le x_j^*$, T^2-statistic is positive and it increases with $|x_j|$. The value of variable x_j contradicts the correlation structure if $x_j \le 0$. On the other hand, a value of $x_j > x_j{}^*$ represents a large positive deviation with respect to the mean, in the direction indicated by the correlation matrix.

Several methods have been also presented to calculate the variables' contributions when latent variable projection methods are used for monitoring purposes. In the case latent variables cannot be associated to a meaningful group of process variables, Nomikos [12] proposed an approach to calculate the contributions of each process variable to the T^2-statistic (called D^2-statistic when projection methods based on Principal Component Analysis (PCA) are applied) instead of to the scores. Westerhuis et al. [13] presented an extended approach of the contribution plots proposed by Nomikos [12] to be applicable also to latent variable models with correlated scores. They proposed to calculate the contribution of the jth variable to the inflated statistic value (c_j) as:

$$D^2 = \mathbf{t}^{\mathbf{T}}\mathbf{S}_{\mathbf{L}}^{-1}\mathbf{t} = \mathbf{t}^{\mathbf{T}}\mathbf{S}_{\mathbf{L}}^{-1}\left[\mathbf{x}^{\mathbf{T}}\mathbf{P}\right]^{\mathbf{T}} \tag{7}$$

$$D^2 = \mathbf{t}^{\mathbf{T}}\mathbf{S}_{\mathbf{L}}^{-1}\sum_{j=1}^{J}\left[x_j\mathbf{p}_j^{\mathbf{T}}\right]^{\mathbf{T}} \tag{8}$$

$$D^2 = \sum_{j=1}^{J}\mathbf{t}^{\mathbf{T}}\mathbf{S}_{\mathbf{L}}^{-1}\left[x_j\mathbf{p}_j^{\mathbf{T}}\right]^{\mathbf{T}} = \sum_{j=1}^{J}c_j \tag{9}$$

where $\mathbf{t}$ and $\mathbf{S_L}$ are the coordinates of $\mathbf{x}$ in the considered latent space and their variance–covariance matrix respectively, $\mathbf{P}$ is the loading matrix with R retained PCs and $\mathbf{p}_j$ represents the j-th column of matrix $\mathbf{P}$. This technique associates no interpretation to the negative values that can result from the calculations.

3. A New approach to estimate variable contributions to T^2-statistic

Assuming that no systematic errors in measurements are present, it is still possible to obtain an observation point whose statistic value exceeds the critical distribution value for a given significance level. In this case, one or more variables in the observation vector do not behave as the observations in the reference population do. The occurrence of an anomalous event is declared, and the question of which variables reveal the faulty state follows. In this work an answer based on the knowledge of the nearest neighbor of the observation point that is in statistical control is provided. This information gives us an idea on how far from an in control allocation the faulty observation is and which directions more explain the occurrence of the anomalous situation.

The problem of finding the NICN can be stated as an optimization problem, whose objective is to determine the coordinates of an alternative point that minimizes a distance measure to the observation point, subject to the constraint that the T^2-statistic value for the NICN is equal to the critical statistic value for a given significance level α, T_C^2. Therefore the optimization problem is formulated as follows

$$\begin{aligned} &Min \quad (\mathbf{x}_{NICN}-\mathbf{x})^{\mathbf{T}}\boldsymbol{\Psi}^{-1}(\mathbf{x}_{NICN}-\mathbf{x}) \\ &s.t. \\ &\qquad \mathbf{x}_{NICN}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{x}_{NICN} = T_C^2 \end{aligned} \tag{10}$$

where $\mathbf{x}_{NICN}$ and $\mathbf{x}$ are the NICN and measurement vectors respectively, $\boldsymbol{\Psi}$ is the matrix that defines the type of distance chosen to measure the proximity to the faulty observation, and $\mathbf{R}$ is the correlation matrix estimated from the reference population. Reader should notice that the calculated nearest neighbor is not an observation, but the closest point in the curve for the critical distance.

If variables are measured in the same units, they are also uncorrelated and have approximately equal variances, it is appropriate to use non-standardized measurement vectors and the inverse of the covariance matrix $\mathbf{S}$ to formulate the optimization problem, and to also select $\boldsymbol{\Psi} = \mathbf{I}$. When different observations are noncommensurable and likely to have very different variances, the use of the unweighted Euclidean metric is inadequate. In this case, the weighting matrix corresponds to the sample correlation matrix, $\boldsymbol{\Psi} = \mathbf{R}$, which is a diagonal matrix only if measurements are uncorrelated.

The Lagrangian function, L, for the optimization problem stated in Eq. (10) is defined as

$$L = (\mathbf{x}_{NICN}-\mathbf{x})^{\mathbf{T}}\boldsymbol{\Psi}^{-1}(\mathbf{x}_{NICN}-\mathbf{x}) - \lambda\left(\mathbf{x}_{NICN}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{x}_{NICN} - T_C^2\right) \tag{11}$$

$$L = \mathbf{x}_{NICN}^{\mathbf{T}}\boldsymbol{\Psi}^{-1}\mathbf{x}_{NICN} + \mathbf{x}^{\mathbf{T}}\boldsymbol{\Psi}^{-1}\mathbf{x} - 2\mathbf{x}^{\mathbf{T}}\boldsymbol{\Psi}^{-1}\mathbf{x}_{NICN} - \lambda\left(\mathbf{x}_{NICN}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{x}_{NICN} - T_C^2\right). \tag{12}$$

A well known result from the theory of nonlinear constrained programming is that any local minimum of problem (10) must satisfy the following necessary conditions of optimality:

$$\frac{\partial L}{\partial \mathbf{x}_{NICN}} = 2\boldsymbol{\Psi}^{-1}\mathbf{x}_{NICN} - 2\lambda\mathbf{R}^{-1}\mathbf{x}_{NICN} - 2\boldsymbol{\Psi}^{-1}\mathbf{x} = 0 \tag{13}$$

$$\frac{\partial L}{\partial \lambda} = \mathbf{x}_{NICN}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{x}_{NICN} - T_C^2 = 0 \tag{14}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers.

If the distance measure to be minimized is chosen as the Mahalanobis distance between the observation and the NICN (i.e. $\boldsymbol{\Psi} = \mathbf{R}$), Eq. (13) can be rewritten as:

$$\frac{\partial L}{\partial \mathbf{x}_{NICN}} = 2(1-\lambda)\mathbf{R}^{-1}\mathbf{x}_{NICN} - 2\mathbf{R}^{-1}\mathbf{x} = 0. \tag{15}$$

Using Eq. (15) the coordinates of the NICN are expressed in terms of the observation vector and the Lagrange multipliers as follows

$$\mathbf{x}_{NICN} = d\ \mathbf{x} \tag{16}$$

where $d = (1-\lambda)^{-1}$. Then d is calculated by replacing Eq. (16) in Eq. (14)

$$d^2\left(\mathbf{x}\mathbf{R}^{-1}\mathbf{x}\right) - T_C^2 = d^2T_{\mathbf{x}}^2 - T_C^2 = 0 \tag{17}$$

$$d = \pm\left(\frac{T_C^2}{T_{\mathbf{x}}^2}\right)^{1/2} \tag{18}$$

where $T_{\mathbf{x}}^2$ is the T^2-statistic for the observation vector $\mathbf{x}$.

Considering that there are only two possible solutions, the comparison of the objective function at both solutions is much easier than evaluating the second order optimality conditions to decide which one is the corresponding NICN.

The solution of the optimization problem for the more general situation, i.e. $\boldsymbol{\Psi} \neq \mathbf{R}$, requires the use of non-linear programming techniques. Nevertheless the computational cost of considering different distances metrics does not increase significantly. The coordinates of the NICN depend on the selection of this metric. To illustrate this issue, let us consider the bidimensional case study represented in Fig. 1. It is assumed that the reference population is centered in [0,0], and there exist two out-of-control observation points, $p_1 = [-1,2]$ and $p_2 = [2,0]$. The figure shows three types of level curves. One of them corresponds to T^2-statistic value curves (full lines), which are ellipses around [0.0]. The others are defined by the Euclidean distance (full lines), which are circles around the two out-of-control observations, and the Mahalanobis distance (dashed lines) from the out of control observations. Also the intersection points between the T^2-statistic value curves and the distance lines are represented and joined by a line. It may be noticed that a straight line

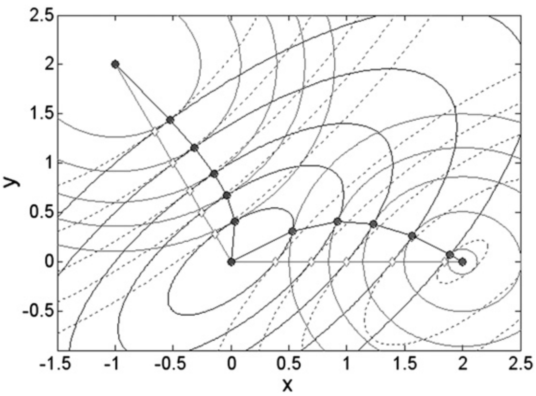

Fig. 1. Different positions of NICN depending on the selection of the distance metric.

is defined by the neighbors when $\Psi = \mathbf{R}$. On the other hand, when the Euclidean distance is considered the intersections follow a non-linear curve.

After calculating the coordinates of the NICN, the influence of each variable to the inflated statistic value is estimated. Since all the variables have been previously standardized to be dimensionless, the distance in which each measured variable should be modified to reach the NICN is considered as the contribution of that variable to the statistic value, i.e.,

$$c_j = abs\left(x_{j,NICN} - x_j\right), \tag{19}$$

where c_j is the contribution of the j-th variable, and $x_{j,NICN}$ and x_j are the j-th elements of vectors $\mathbf{x}$ and $\mathbf{x}_{NICN}$, respectively. Therefore the resulting movements in each direction can be applied in a similar way as classically used in contribution plots. The directions whose changes are greater than their threshold values are considered as suspicious variables.

To determine the threshold value, a data-driven procedure is implemented based on the information contained in the reference population. For each variable, its contribution to the inflated statistic for a set of simulated fault cases, in which that variable is not faulty, is calculated. Then the empirical cumulative distribution function of that variable contribution is obtained. The control limit is selected as the variable-contribution value for which the cumulative probability is $(1-\alpha)$, where α is the probability of wrongly identifying a variable as faulty and is set in the range [0.05–0.1].

4. Case studies

In this section, the results of a performance comparison among the new strategy and other methodologies are presented. Variable contributions to the T^2-statistic are obtained by applying the NICN strategy, and compared with those estimated by using the OSS (Alvarez et al. [6]), and the ones calculated by using Eqs. (7) to (9) due to Westerhuis et al. [13].

Regarding the NICN approach, the Mahalanobis and Euclidean distances are selected as objective functions of the optimization problem. As it is expected, the identification performance of the procedure that uses the Mahalanobis distance is better than the other, because variables are correlated for the analyzed case studies. Therefore, only the results obtained using the Mahalanobis distance are included in this work.

With respect to the OSS and PCA strategies, a threshold value τ_j for each variable is calculated using the information of the mean $(\bar{c}_j)$ and standard deviation (s_j) of the j-th variable contribution to the corresponding statistics for the reference population samples, as follows

$$\tau_j = \bar{c}_j + \beta_j s_j. \tag{20}$$

Parameters β_j are selected such that the best identification results of both techniques are reported for the analyzed case studies.

Table 1
Numerical example. Data reported by De Maesschalck et al. [14].

Observation	x_1	x_2	x_3	x_4
1	4.00	3.00	1.00	2.00
2	5.00	4.00	2.00	3.50
3	8.00	7.00	3.00	4.00
4	8.00	6.00	5.00	4.00
5	9.00	7.00	2.00	3.00
6	6.00	3.00	5.00	3.00
7	6.00	5.00	3.00	2.50
8	10.00	8.00	2.00	3.00
9	2.00	3.00	1.50	3.40
10	4.00	4.00	3.00	3.00
11	6.00	6.00	6.00	4.00
12	6.50	4.50	0.00	2.00
13	9.00	8.00	5.00	5.00
14	4.00	5.00	1.00	1.00
15	4.00	6.00	3.00	5.00
16	6.00	7.00	2.00	4.00
17	2.50	4.50	6.00	4.00
18	5.00	5.50	8.00	3.00
19	7.00	5.50	1.00	2.50
20	8.00	5.00	3.00	3.00

The significance level for all statistical tests used for detection and identification purposes is fixed at 0.05.

Application examples extracted from the literature are selected as case studies. Even though sample populations must be regarded as small, these are kept as originally published for the sake of both, simplicity and reproducibility.

4.1. Case Study I

Let us consider the data set presented by De Maesschalck et al. [14] as a reference population. It is constituted by 20 observations of four variables which are reported in Table 1. The corresponding mean vector is $\bar{x} = [6 \quad 5.35 \quad 3.125 \quad 3.245]$. In addition, seven test observations included in Table 2 are used to show how the three aforementioned identification strategies interpret their Hotelling-statistic's values. The T^2-statistic value for each test observation and the critical statistic value are also reported in the same table.

The pair of measurements $TEST_1/TEST_2$ has a deviation of the same magnitude but different sign for variable 1. Therefore these test measurements have the same Euclidean distance from the mean vector and also the same statistic value. It is independent of the deviation sign because the three remaining variables are at their mean values. The same behavior is noticed for the pair of measurements $TEST_3/TEST_4$ even though two variables deviate with respect to their means. In this case the deviations of variables 1 and 2 for $TEST_4$ have the same magnitude but different sign with respect to $TEST_3$. Regarding $TEST_5$ to $TEST_7$, variable 3 presents a significant deviation with respect to the mean in $TEST_5$, the same situation occurs for variables 2 and 4 in $TEST_6$, and for variables 1 and 3 in $TEST_7$, respectively.

Table 2
Test observations for Case Study I.

Test number	Observation	T^2
$TEST_1$	[−1.099 5.350 3.125 3.245]	**24.029**
$TEST_2$	[13.09 5.350 3.125 3.245]	**24.029**
$TEST_3$	[−1.099 6.124 3.125 3.245]	**30.957**
$TEST_4$	[13.09 4.576 3.125 3.245]	**30.957**
$TEST_5$	[7.996 6.991 12.88 5.003]	**23.621**
$TEST_6$	[3.782 9.376 3.498 6.998]	**31.299**
$TEST_7$	[0.454 6.898 8.317 44.228]	**29.035**

$T^2_{C,0.05} = 14.997$. Values in bold and underline indicate that the corresponding statistic detects a fault.

Table 3
NICN variable contributions for Case Study I.

Observation	$c_1^{T^2}$	$c_2^{T^2}$	$c_3^{T^2}$	$c_4^{T^2}$
$TEST_1$, $TEST_2$	**0.6720**	0.0000	0.0000	0.0000
$TEST_3$, $TEST_4$	**0.9727**	0.1520	0.0000	0.0000
$TEST_5$	0.1829	0.2154	**0.9551**	0.3599
$TEST_6$	0.3077	**0.8001**	0.0554	**1.1640**
$TEST_7$	**0.7033**	0.2813	**0.7033**	0.2787

Values in bold and underline indicate that the corresponding statistic detects a fault.

In Tables 3 and 4, the variable contributions to the T^2-statistic calculated using the NICN and OSS approaches are reported. For $TEST_6$ and $TEST_7$, OSS strategy gives ambiguous identifications [7]. These arise when the original cause of the fault is pointed out as a suspicious variable along with others, which influence can be explained based on the engineering knowledge about the process. In contrast, the NICN methodology provides precise identifications for all the test observations, that is, only the faulty variables are indicated as the suspicious ones.

Furthermore, the decomposition method proposed by Mason et al. [4], which also works in the original measurement space, is applied to Case Study I. When the test significance level α is set at 0.05, the technique gives an ambiguous identification associated to variable 1 for $TEST_6$, and all other identifications are precise. In light of that methodology, all tests are run using lower significance levels ($\alpha=0.01$ and $\alpha=0.001$). No suspicious variables are identified for $\alpha=0.001$, i.e., all identifications are void. Using $\alpha=0.01$, the identifications are precise for $TEST_5$, ambiguous for $Test_6$ and $Test_7$ and void for the remaining tests. Therefore, $\alpha=0.05$ provides satisfactory results.

Since the strategy developed by Westerhuis et al. [13] uses a latent variable model, a PCA model of the reference data was performed considering 2 and 3 retained PCs, R, reaching a variance reconstruction of 82.6% and 95%, respectively. Table 5 shows the corresponding values for D^2 and SPE statistics. Values in bold and underline indicate that the corresponding statistic detects a fault.

Detection capabilities of Hotelling's statistic are strongly affected by the dimension reduction. Only $TEST_5$ is detected as a faulty observation by the D^2-statistic for $R=2$. If R increases to 3, both $TEST_5$ and $TEST_6$ are indicated as faulty measurements. Consequently, only values for those tests will be analyzed for identification performance comparisons. Table 6 shows the values of the contributions to D^2-statistic calculated using Eqs. (7) to (9). A precise fault identification for each test is obtained.

To complete the analysis, the performance of the SPE-statistic is also evaluated. It exceeds the critical value for all tests, except the $TEST_5$ for $R=3$. Therefore the PCA procedure detects the out of control state for all the test observations, since it uses both the D^2 and SPE statistics. Regarding its identification performance, variable contributions to this statistic are calculated using the methodology proposed by Westerhuis et al. [13]. Table 6 shows that all contributions are above their control limits for $R=3$, and many of them are also greater than their threshold values for $R=2$. Therefore it is difficult to determine the right set of faulty observations. Furthermore, the identification is incorrect for $TEST_7$ and $R=2$, because one of the faulty variables is not highlighted as suspicious [7].

Table 4
OSS variable contributions for Case Study I ($\beta=3$).

Observation	$c_1^{T^2}$	$c_2^{T^2}$	$c_3^{T^2}$	$c_4^{T^2}$
$TEST_1$, $TEST_2$	**24.03**	0.000	0.000	0.000
$TEST_3$, $TEST_4$	**27.14**	3.821	0.000	0.000
$TEST_5$	1.065	−0.167	**24.23**	−1.511
$TEST_6$	**5.411**	**14.19**	−0.401	**12.09**
$TEST_7$	**17.95**	**6.896**	**6.541**	−2.357

Values in bold and underline indicate that the corresponding statistic detects a fault.

Table 5
D and SPE statistics when PCA is applied for Case Study I.

	D	SPE	R
$TEST_1$, $TEST_2$	5.75	**3.67**	3
$TEST_3$, $TEST_4$	5.17	**5.18**	3
$TEST_5$	**23.62**	0.01	3
$TEST_6$	**24.28**	**1.40**	3
$TEST_7$	7.79	**4.27**	3
$TEST_1$, $TEST_2$	3.46	**3.17**	2
$TEST_3$, $TEST_4$	2.66	**4.52**	2
$TEST_5$	**13.86**	**2.52**	2
$TEST_6$	6.72	**8.95**	2
$TEST_7$	7.77	**4.28**	2
$D_{C,0.05}=11.2550$	$SPE_{C,0.05}=0.8100$		3
$D_{C,0.05}=7.8793$	$SPE_{C,0.05}=2.3866$		2

Values in bold and underline indicate that the corresponding statistic detects a fault.

4.2. Case Study II

The second case study is a tubular reactor where the reaction $A+B\rightarrow 3C$ takes place. The set of measured variables is composed by ten observations: the inlet composition of A, B and C compounds, inlet reactor and refrigerant temperatures, inlet flowrate, reactor temperature at axial positions corresponding to 1/3 and 2/3 of the reactor length, outlet reactor temperature and outlet composition of C compound, which are identified as variables 1 to 10, respectively. The reference population is formed by thirty seven observations whose mean vector and covariance matrix are shown in Table 7.

Seven additional tests are considered to perform the same comparisons as in the previous case. The first six runs represent the presence of faults associated to the input variables 1 to 6, respectively. For Run 7, faults are simulated for variables 4 and 5. The standardized observation vector and the T^2-statistic value of each run are provided in Table 8.

In Tables 9 and 10, the variable contributions to the T^2-statistic calculated using the NICN and OSS approaches are reported. For this case study, only the faulty variables are identified as suspicious using the NICN procedure, except for Run 1. In this test, the simulated change in the inlet composition of A leads the outlet composition of C to be out of the normal region. From Table 10, it can be seen that the OSS strategy gives a precise identification for Run 5, ambiguous identifications for Runs 1, 2, 3 and 6, and incorrect identifications for Runs 4 and 7.

For this case study, the methodology presented by Mason et al. [4] gives ambiguous identifications for five runs (Runs 1–4, Run 6), a precise identification for Run 5 and an incorrect identification for Run 7, when it is run using $\alpha=0.05$. For this α value, the ambiguous identifications signal an average of four suspicious variables besides the faulty one. The reduction of α value to 0.01 has no significant effect

Table 6
Variable contributions to D^2 and SPE statistics for Case Study I ($\beta=2$).

Observation	c_1^D	c_2^D	c_3^D	c_4^D	D	R
$TEST_5$	0.9895	−0.0597	**24.283**	−1.5957	**23.62**	3
$TEST_6$	1.8727	**3.1415**	−0.4813	**19.750**	**24.28**	3
$TEST_5$	−0.3388	0.4311	**10.241**	3.5294	**13.86**	2
	c_1^{SPE}	c_2^{SPE}	c_3^{SPE}	c_4^{SPE}	SPE	
$TEST_1$, $TEST_2$	**1.3195**	**1.9035**	**0.0210**	**0.4317**	**3.6757**	3
$TEST_3$, $TEST_4$	**1.8612**	**2.6850**	**0.0296**	**0.6090**	**5.1848**	3
$TEST_6$	**0.5061**	**0.7301**	**0.0081**	**0.1656**	**1.4098**	3
$TEST_7$	**1.5335**	**2.2123**	**0.0244**	**0.5018**	**4.2720**	3
$TEST_1$, $TEST_2$	**2.2580**	**2.2223**	0.3267	0.0014	**4.8084**	2
$TEST_3$, $TEST_4$	**3.0122**	**3.0804**	0.3359	0.0027	**6.4313**	2
$TEST_5$	**0.5595**	0.0623	**2.1838**	**2.0269**	**4.8326**	2
$TEST_6$	**2.8639**	1.3508	**3.5944**	**2.2990**	**10.108**	2
$TEST_7$	**1.4511**	**2.1809**	0.0504	0.5998	**4.2822**	2

Values in bold and underline indicate that the corresponding statistic detects a fault.

Table 7
Tubular reactor example.

	1	2	3	4	5	6	7	8	9	10
Mean	0.2108	4.0603	0.0019	624.32	624.79	0.0997	634.61	628.89	627.00	0.5718
Cov	1	2	3	4	5	6	7	8	9	10
1	0.0001	0.0002	0.0000	−0.0478	−0.0002	0.0000	0.0051	0.0023	0.0004	0.0002
2	0.0002	0.0202	0.0000	−0.7030	−0.0078	0.0001	0.0328	0.0065	−0.0135	0.0011
3	0.0000	0.0000	0.0000	−0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	−0.0478	−0.7030	−0.0004	111.79	2.5101	0.0003	−7.9096	−1.2522	1.4182	−0.1076
5	−0.0002	−0.0078	0.0000	2.5101	1.0006	0.0003	0.6604	0.6321	0.8944	0.0035
6	0.0000	0.0001	0.0000	0.0003	0.0003	0.0000	0.0003	0.0002	0.0003	0.0000
7	0.0051	0.0328	0.0000	−7.9096	0.6604	0.0003	1.3926	0.7216	0.6898	0.0144
8	0.0023	0.0065	0.0000	−1.2522	0.6321	0.0002	0.7216	0.7668	0.6041	0.0081
9	0.0004	−0.0135	0.0000	1.4182	0.8944	0.0003	0.6898	0.6041	0.8229	0.0041
10	0.0002	0.0011	0.0000	−0.1076	0.0035	0.0000	0.0144	0.0081	0.0041	0.0005

Table 8
Standardized runs and T^2-statistic for Case Study II.

Run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	T^2
R1	−3.977	0.441	−1.177	−0.839	1.542	−0.107	−0.356	0.278	0.190	−4.273	**300.7**
R2	−0.189	4.299	−1.367	−0.891	−1.037	−1.713	−0.017	−0.364	−0.374	0.690	**168.9**
R3	1.018	−1.203	−5.409	0.168	0.792	−0.949	0.426	0.625	0.425	0.828	**76.03**
R4	−0.494	0.499	−0.095	−1.700	−1.818	0.684	1.122	0.531	0.126	−0.579	**372.9**
R5	−0.494	0.499	−0.095	0.501	3.281	0.696	−0.164	0.375	0.371	−0.168	**940.5**
R6	−1.066	1.487	0.542	−0.048	−0.328	3.982	0.562	0.367	0.189	−0.299	**105.9**
R7	0.003	0.499	−0.095	−1.500	−2.633	0.694	1.038	0.416	0.029	−0.053	**748.9**

$T^2_{C,0.05}=30.2$. Values in bold and underline indicate that the corresponding statistic detects a fault.

on the identification categories, but the ambiguous identifications signal an average of two suspicious variables besides the faulty one in this case. For $\alpha=0.001$, the identification performance is worse than that for $\alpha=0.01$.

A PCA of the same data has been carried out giving a total variance reconstruction of 74.5% and 64% for $R=3$ and $R=2$, respectively. Cattell's criterion [9] has been used to choose the number of retained P.C.s. No runs are detected as faulty observations by the D^2-statistic for $R=2$. When the PCA model built for $R=3$ is used to evaluate the test observations, only *R1* and *R3* are pointed out as out-of-control observations by the D^2 statistic. The variables' contributions are shown in Table 11. It can be seen that the main contributions correspond to the actual simulated deviations but the identifications are ambiguous. The *SPE*-statistic detects that the process is out of control for all the test runs, but no identifications are precise, as it is reported in Table 11.

Table 9
NICN variable contributions for Case Study II.

Run	$c_1^{T^2}$	$c_2^{T^2}$	$c_3^{T^2}$	$c_4^{T^2}$	$c_5^{T^2}$	$c_6^{T^2}$	$c_7^{T^2}$	$c_8^{T^2}$	$c_9^{T^2}$	$c_{10}^{T^2}$
R1	**2.717**	0.302	0.804	0.573	1.053	0.073	0.243	0.189	0.129	**2.919**
R2	0.109	**2.482**	0.789	0.514	0.599	0.677	0.010	0.210	0.216	0.398
R3	0.376	0.445	**2.001**	0.062	0.293	0.351	0.158	0.231	0.157	0.306
R4	0.354	0.357	0.068	**1.216**	1.301	0.497	0.803	0.379	0.090	0.415
R5	0.406	0.410	0.078	0.410	**2.693**	0.569	0.135	0.308	0.305	0.138
R6	0.497	0.694	0.253	0.022	0.153	**1.857**	0.262	0.171	0.088	0.139
R7	0.002	0.399	0.076	**1.199**	**2.105**	0.555	0.829	0.332	0.024	0.043

Values in bold and underline indicate that the corresponding statistic detects a fault.

5. Conclusions

A novel method to estimate the variable contributions to the T^2-statistic is presented. Given a measured point whose T^2-statistic value exceeds the critical value T^2_C, the contribution of each variable is determined in terms of the minimum distance between the measured point and its closer neighbor whose T^2-statistic value is equal to T^2_C. The variables that help to reveal the faulty state are those whose contributions exceed their control limits. A data driven technique is devised to determine those threshold values. Results have shown a good performance when this technique is applied in the original variable space.

The calculation of the T^2-statistic is based on the correlation matrix that should be nonsingular to be inverted. Problems in applying the proposed technique may arise when the number of variables is large in comparison with the number of samples. In this case the correlation matrix becomes ill-conditioned, and it is necessary to apply a regularization procedure to tackle the problem. Also the presence of collinear variables originates difficulties with the inversion of the

Table 10
OSS variable contributions for Case Study II ($\beta=3$).

Run	$c_1^{T^2}$	$c_2^{T^2}$	$c_3^{T^2}$	$c_4^{T^2}$	$c_5^{T^2}$	$c_6^{T^2}$	$c_7^{T^2}$	$c_8^{T^2}$	$c_9^{T^2}$	$c_{10}^{T^2}$
R1	**34.19**	−11.41	2.819	**49.39**	**262.2**	−0.493	13.97	2.207	−27.22	−24.97
R2	−0.895	**108.0**	3.479	−30.54	**134.1**	3.892	−0.337	0.946	−43.54	−6.167
R3	−18.68	**13.83**	**43.05**	−1.213	**48.97**	−0.705	4.017	2.885	−30.64	14.51
R4	−2.308	**14.64**	0.116	−71.67	**378.4**	−0.752	**23.52**	−0.915	24.24	7.640
R5	12.18	−25.05	0.144	−37.73	**1101**	3.688	5.274	2.827	−117.0	−5.231
R6	19.84	**17.70**	−1.857	−1.741	26.51	**15.45**	**23.12**	0.563	8.928	−2.529
R7	0.036	**21.67**	0.078	−98.58	**785.9**	−1.812	**34.04**	−1.690	8.115	1.121

Values in bold and underline indicate that the corresponding statistic detects a fault.

Table 11
Variable contributions to D^2 and SPE statistics for Case Study II ($\beta = 3$).

Run	c_1^D	c_2^D	c_3^D	c_4^D	c_5^D	c_6^D	c_7^D	c_8^D	c_9^D	c_{10}^D	D	R
R1	**6.391**	0.185	−1.669	−0.215	0.670	0.062	0.036	0.040	0.060	**5.863**	**11.42**	3
R3	1.622	**2.121**	**15.82**	0.165	0.108	−0.719	−0.133	−0.042	0.108	1.086	**20.13**	3
	c_1^{SPE}	c_2^{SPE}	c_3^{SPE}	c_4^{SPE}	c_5^{SPE}	c_6^{SPE}	c_7^{SPE}	c_8^{SPE}	c_9^{SPE}	c_{10}^{SPE}	SPE	
R1	**1.005**	0.593	**5.478**	**5.262**	**0.665**	0.710	0.314	0.184	0.100	**2.569**	**16.88**	3
R2	**0.497**	**9.702**	**4.889**	0.179	0.012	1.296	0.016	0.027	**0.550**	0.077	**17.25**	3
R3	**0.391**	0.897	**4.388**	**1.340**	0.109	3.202	0.516	0.263	0.026	0.295	**11.43**	3
R4	**0.555**	0.086	0.672	0.759	**1.798**	0.529	**0.771**	0.417	**0.343**	**0.667**	**6.599**	3
R5	0.001	0.843	0.084	0.015	**3.864**	0.506	0.441	0.311	**0.849**	0.008	**6.923**	3
R6	**1.146**	0.925	0.034	0.173	0.202	**16.22**	0.023	0.008	0.004	0.145	**18.88**	3
R7	**0.429**	0.110	0.301	0.215	**3.248**	0.380	**0.734**	0.548	**0.667**	0.375	**7.008**	3
R1	**4.371**	**3.242**	0.262	**7.844**	**0.748**	0.126	0.594	0.299	0.119	**6.215**	**23.82**	2
R2	1.188	**12.11**	2.446	0.059	0.016	1.718	0.003	0.042	**0.535**	0.001	**18.11**	2
R3	**1.744**	0.802	**28.80**	0.063	0.058	0.844	0.117	0.091	0.012	1.099	**33.63**	2
R4	1.162	0.001	0.068	1.053	**1.758**	0.335	**0.888**	0.465	**0.333**	1.186	**7.248**	2
R5	0.014	1.014	0.018	0.028	**3.880**	0.449	0.417	0.300	**0.854**	0.000	**6.975**	2
R6	**1.993**	1.651	0.150	0.066	0.189	**15.01**	0.048	0.016	0.003	0.435	**19.56**	2
R7	0.645	0.037	0.090	0.283	**3.224**	0.303	**0.784**	0.572	**0.661**	0.537	**7.136**	2
	$D_{C,0.05} = 9.404$			$SPE_{C,0.05} = 5.378$								3
	$D_{C,0.05} = 6.903$			$SPE_{C,0.05} = 6.707$								2

Values in bold and underline indicate that the corresponding statistic detects a fault.

correlation matrix. These cases should also be detected and then, an adjustment for the presence of collinearities is required using the techniques described by Mason and Young [2].

Some differences appear between the identification capabilities of OSS and NICN strategies. They are due to the fact that all the directions are modified when the NICN approach is applied in contrast to the "fixed curve" approach given by the OSS, which can produce high negative values for some contributions.

The NICN approach works in the original variable space and therefore employs only the T^2-statistic. This avoids the possible loss of information originated by the projection into an incorrectly dimensioned P.C. space, which may lead to detection faults (Type II Error), and false alarms (Type I Error), as it was shown in previous works [6].

Acknowledgments

The authors wish to thank CONICET (Consejo Nacional de Investigaciones Científicas y Tecnológicas, Argentina), UNS (Universidad Nacional del Sur, Argentina) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Argentina) for the financial support.

References

[1] H. Hotelling, The generalization of student's ratio, The Annals of Mathematical Statistics 2 (1931) 360–378.
[2] R.L. Mason, J.C. Young, Multivariate statistical process control with industrial applications, ASA SIAM Series on Statistics and Applied Probability, 2002. Philadelphia; Pennsylvania.
[3] P.J. Bickel, E. Levina, Regularized estimation of large covariance matrix, The Annals of Statistics 38 (2008) 199–227.
[4] R.L. Mason, N.D. Tracy, J.C. Young, Decomposition of T2 for multivariate control chart interpretation, Journal of Quality Technology 27 (1995) 99–108.
[5] R.L. Mason, N.D. Tracy, J.C. Young, A practical approach for interpreting multivariate T2 control chart signals, Journal of Quality Technology 29 (1997) 396–406.
[6] C.R. Alvarez, A. Brandolin, M.C. Sánchez, On the variable contributions to the D-statistic, Chemometrics and Intelligent Laboratory Systems 88 (2007) 189–196.
[7] C.R. Alvarez, A. Brandolin, M.C. Sánchez, Batch process monitoring in the original measurement's space, Journal of Process Control 20 (2010) 716–725.
[8] M.C. Sánchez, C.R. Alvarez, A. Brandolin, A MSPC procedure for bias identification in steady-state processes, AIChE Journal 54 (2008) 2082–2088.
[9] J.E. Jackson, A User's Guide To Principal Components, John Wiley & Sons, Inc., United States of America, 1991.
[10] P. Miller, R.E. Swanson, C.F. Heckler, Contribution plots: the missing link in multivariative quality control, Applied Mathematics and Computer Science 8 (1999) 775–792.
[11] J.F. Macgregor, C. Jaeckle, C. Kiparissides, M. Koutoudi, Process monitoring and diagnosis by multiblock PLS methods, AICHE Journal 40 (1994) 826–838.
[12] P. Nomikos, J.F. Macgregor, Monitoring of batch process using multiway principal component analysis, AICHE Journal 40 (1994) 1361–1375.
[13] 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.
[14] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemometrics and Intelligent Laboratory Systems 50 (2000) 1–18.