Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 88 (2007) 189 – 196
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

On the variable contributions to the *D*-statistic

Carlos Rodrigo Alvarez, Adriana Brandolin, Mabel Cristina Sánchez *

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

Received 20 October 2006; received in revised form 30 April 2007; accepted 30 April 2007
Available online 10 May 2007

Abstract

The *D*-statistic is widely used in Statistical process control to reliably detect the out of control status, but by itself it offers no assistance as fault identification tool. Some strategies, that work in the original or in the latent variable space, have been proposed to show the contribution of each process variable to the calculated statistic. Nevertheless it is still an open research subject.

In this work, a straightforward strategy to decompose the *D*-statistic as a unique sum of each variable contribution is presented, that is applied in the space of the original variables. Also an explanation is provided regarding the physical meaning of the negative contributions to the statistic. The results of the proposed strategy are compared with those obtained in the latent variable space using other methods. As the new strategy works in the original variable space, the selection of an appropriate method to calculate the number of retained latent variables which reduce the lost of significant information, is avoided.

Keywords: *D*-statistic; Contribution plots; Statistical process control; Fault identification; Process monitoring

1. Introduction

Statistical process monitoring involves three activities: detection of the out-of-control status, identification of the variable(s) that signal this condition, and diagnosis of the source cause for the abnormal behaviour. Monitoring focuses on the detection and identification activities, while diagnosis provides the information for determining the corrective action. The *D*-statistic can detect the out of control status reliably but offers no assistance in the identification of the variables responsible for this status. Some strategies have been proposed to assign variable-contribution values to the *D*-statistic taking into account the multivariate nature of process data.

Regarding the methods that transform the data from the original **X**-space to a latent variable-space, Jackson [1] proposed the decomposition of the statistic into a sum of principal components. If they represent a meaningful grouping of variables, the identification of out-of-control signals is readily apparent [2]. However in many examples it is difficult to associate a meaning to a principal component and the characteristics associated with out-of-control signals cannot be determined.

If the *D*-statistic value is too large, which indicates that the process is out of control, the scores of a new observation will be at a large Mahalanobis distance from the center of the empirical model based on Principal Component Analysis (PCA). Miller et al. [3] and MacGregor et al. [4] proposed to evaluate the contributions of each process variable to the scores that are outside of their confidence limits. Furthermore Nomikos [5] presented an approach to calculate the contributions of each process variable to the *D*-statistic instead of to the scores, when latent variables cannot be associated to a meaningful group of process variables.

Westerhuis et al. [6] 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 which behaviour are different with respect to those contained in the reference data set.

Another approach for calculating variable contributions to the *D*-statistic is carried out in the original **X**-space. Mason et al. [7] provided a technique to decompose the *D*-statistic value into a sum of N independent parts, where N is the number of measured variables. The first term is calculated squaring a univariate t statistic for one variable. The j-th term ($j=2,\ldots, N$) of the sum is another component of the measurement vector adjusted by the estimates of the mean and standard deviation of its conditional probability distribution given the ($j-1$) previously incorporated variables. As

* Corresponding author. Tel.: +54 291 486 1700; fax: +54 291 486 1600.
E-mail address: msanchez@plapiqui.edu.ar (M.C. Sánchez).

0169-7439/$ - see front matter
doi:10.1016/j.chemolab.2007.04.007

there exists no fixed order for incorporating measurements, the decomposition of D-statistic into N independent components is not unique. Certainly N! different non-independent partitions are possible. The authors mention that frequently, for each partition the main interest is focused only in two terms: the one corresponding to the unadjusted contribution of a single selected variable and, the term containing the adjusted contributions of one of the variables after the adjustment of the $(N-1)$ remaining ones. Nevertheless if the inspection of this reduced set of terms is not enough to come to a conclusion, all significant conditional terms should be compared to a critical value, increasing the complexity of the identification of the source fault. Later on, Mason et al. [8] introduced a sequential computational scheme that reduces the number of required terms that need to be computed in a decomposition.

In this work a new and straightforward strategy to evaluate variable contributions to the D-statistic is presented, that is carried out in the original **X**-space. Also an explanation is provided regarding the physical meaning of the negative contributions to the statistic. The results of the proposed strategy are compared to those obtained by Westerhuis et al. [6] in the latent variable space.

2. The *D*-statistic

Let $\mathbf{x}$ represent an N dimensional vector of measurements made on a process at time t. When process is in control, it is assumed that the $\mathbf{x}$ vectors are independent and follow a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The population parameters are estimated using the sample mean vector $(\overline{\mathbf{x}})$ and the sample covariance matrix $(\mathbf{S})$ from a reference sample containing I observations.

A multivariate control chart for the process is based on the D-statistic, which has the form

$$D = (\mathbf{x} - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) \tag{1}$$

and follows a $[N(I^2-1)/(I^2-IN)]F_{N,I-N,\alpha}$ distribution where, $F_{N,I-N,\alpha}$ is the value of the F distribution for a level of significance α, with N and $(I-N)$ degrees of freedom.

As it is shown in Eq. (1), the statistic has a quadratic form and its value is always equal or greater than zero considering that the covariance matrix is positive semi-definite. The minimum value of the statistic is achieved when $\mathbf{x}=\overline{\mathbf{x}}$. In the case that only one variable deviates from its mean value, for instance, $x_j=\overline{x}_j$, $j=2$, …, N and, $x_1 \neq \overline{x}_1$, a positive value of the D-statistic might be, a priori, associated with the deviation of x_1 with respect to $\overline{x}_1$, and considered as the contribution of this variable to the statistic.

Furthermore notice that if two vectors of measurements are at the same Euclidean distance from $\overline{\mathbf{x}}$, but one of them deviates in contradiction with the behaviour indicated by the covariance matrix structure, the D-statistic value increases more for this one, as it is exemplified in the next section.

3. Decomposition of the *D*-statistic

Let us consider the simple case in which a population represented by two variables (x_1 and x_2) is monitored. The mean vector is $\overline{\mathbf{x}}=[\overline{x}_1\ \overline{x}_2]$ and correlation matrix $\mathbf{S}$ is

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \rho_{12}s_1s_2 \\ \rho_{12}s_1s_2 & s_2^2 \end{bmatrix} \tag{2}$$

where s_1 and s_2 are the standard deviation for x_1 and x_2 respectively and, ρ_{12} is the correlation factor between x_1 and x_2.

Eq. (1) for the D-statistic can be reformulated in terms of the variable deviations as follows:

$$D = \frac{1}{(1-\rho_{12})^2 s_1 s_2}\left(\frac{s_2}{s_1}(x_1-\overline{x}_1)^2 + \frac{s_1}{s_2}(x_2-\overline{x}_2)^2 - 2\rho_{12}(\mathbf{x}_1-\overline{\mathbf{x}}_1)(x_2-\overline{x}_2)\right) \tag{3}$$

In accordance with the correlation matrix, for given magnitudes of variable deviations, if $\rho_{12}>0$ the smallest value of D will be achieved when both variable deviations present the same sign. On the other hand, if $\rho_{12}<0$ the smallest value occurs when the variable deviations are of opposite sign.

The D-statistic for a new observation is zero if $x_1=\overline{x}_1$ and $x_2=\overline{x}_2$. If $x_1 \neq \overline{x}_1$, the minimum value of the statistic given $\mathbf{x}_1$ is not achieved for $x_2=\overline{x}_2$. In Fig. 1 a plot of D in terms of x_1 and x_2 is presented. The value of variable x_2 for which D is minimum given x_1 is indicated as x_2^{*,x_1}. As can be seen in this figure, x_2^{*,x_1} depends on the selected value for x_1.

For the monitoring of N variables, the D-statistic can be formulated as

$$D = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}x_ix_j, \tag{4}$$

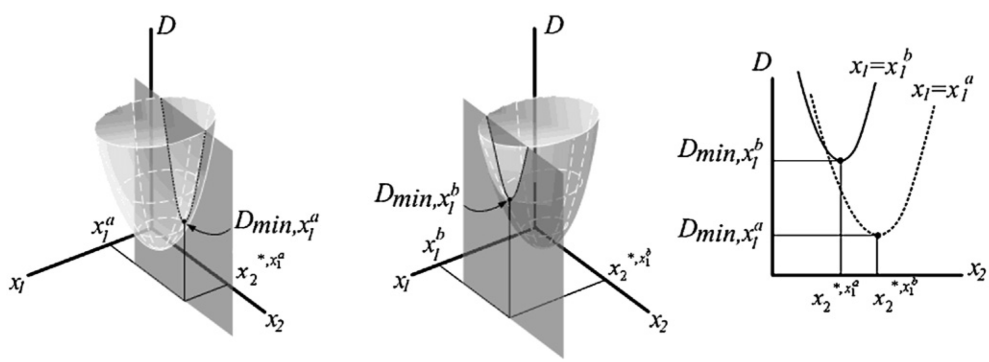

Fig. 1. $D(x_1, x_2)$ surfaces and $D(x_2$, parameter: $x_1)$.

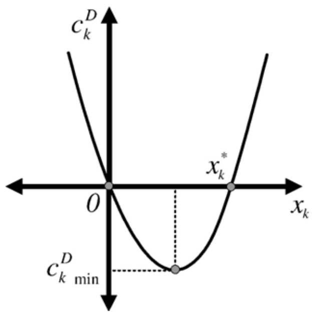

Fig. 2. Contribution of the *k*-variable to *D*.

where a_{ij} are the elements of the inverse of the covariance matrix ($\mathbf{A}=\mathbf{S}^{-1}$) and, for the sake of simplicity it is considered that

$$x_k = x_k^{\text{measured}} - \overline{x}_k. \tag{5}$$

for $k=1,\ldots,N$.

Let us reformulate the *D*-statistic in terms of the variable x_k as follows and, call it D_k

$$D = D_k = a_{k,k}x_k^2 + \left(2\sum_{\substack{j=1\\j\neq k}}^{N} a_{k,j}x_j\right)x_k + \sum_{\substack{i=1\\i\neq k}}^{N}\sum_{\substack{j=1\\j\neq k}}^{N} a_{i,j}x_ix_j. \tag{6}$$

The value of x_k (x_k^*) for which *D* is minimum given the $(N-1)$ values of the remaining variables (D_k^{MIN}) is calculated using the following partial derivative equation

$$\left.\frac{\partial D_k}{\partial x_k}\right|_{x_i=cte;i=1..N;i\neq k} = \left.\frac{\partial D}{\partial x_k}\right|_{x_i=cte;i=1..N;i\neq k} = 2a_{k,k}x_k + \left(2\sum_{\substack{j=1\\j\neq k}}^{N} a_{k,j}x_j\right). \tag{7}$$

The resulting formulas for x_k^* and D_k^{MIN} are the following

$$x_k^* = -\frac{\sum_{\substack{j=1\\j\neq k}}^{N} a_{k,j}x_j}{a_{k,k}}, \tag{8}$$

$$D_k^{MIN} = -\frac{\left(\sum_{\substack{j=1\\j\neq k}}^{N} a_{k,j}x_j\right)^2}{a_{k,k}} + \sum_{\substack{i=1\\i\neq k}}^{N}\sum_{\substack{j=1\\j\neq k}}^{N} a_{i,j}x_ix_j = -a_{k,k}x_k^{*2} + \sum_{\substack{i=1\\i\neq k}}^{N}\sum_{\substack{j=1\\j\neq k}}^{N} a_{i,j}x_ix_j, \tag{9}$$

and, the difference between the value of the *D*-statistic for the observed value for variable x_k and D_k^{MIN} *is*

$$D - D_k^{MIN} = a_{k,k}\left(x_k - x_k^*\right)^2. \tag{10}$$

It would be valuable to use this difference to measure the contribution of the x_k variable to the statistic value, but it should be notice that some problems may arise. For example, if all variables except x_1 are at their mean values and, variable x_j is strongly correlated with x_1, the contribution measure of x_j estimated using Eq.(10) will be large, even if x_j remains at its mean value. To avoid these problems, a new formulation is developed in which Eq. (10) is included.

The sum of the D_k^{MIN} *for* $k=1,\ldots, N$ *is*

$$\sum_{k=1}^{N} D_k^{MIN} = \sum_{k=1}^{N}\left(-a_{k,k}x_k^{*2} + \sum_{\substack{i=1\\i\neq k}}^{N}\sum_{\substack{j=1\\j\neq k}}^{N} a_{i,j}x_ix_j\right) = \sum_{k=1}^{N}\left(-a_{k,k}x_k^{*2}\right) + \sum_{k=1}^{N}\left(\sum_{\substack{i=1\\i\neq k}}^{N}\sum_{\substack{j=1\\j\neq k}}^{N} a_{i,j}x_ix_j\right), \tag{11}$$

and considering that $a_{i,j}=a_{j,i}$, Eq. (11) is re-written as follows

$$\sum_{k=1}^{N} D_k^{MIN} = (N-2)D + \sum_{k=1}^{N} a_{k,k}\left(x_k^2 - x_k^{*2}\right). \tag{12}$$

The *D*-statistic is obtained from Eq. (12). It is evaluated as the sum of the contributions of each variable, c_k^D $(k=1,\ldots, N)$

$$D = \sum_{k=1}^{N} \frac{a_{k,k}}{2}\left[\left(x_k - x_k^*\right)^2 + \left(x_k^2 - x_k^{*2}\right)\right] = \sum_{k=1}^{N} a_{k,k}\left(x_k^2 - x_k^*x_k\right) = \sum_{k=1}^{N} c_k^D \tag{13}$$

The contribution c_k^D is represented in Fig. 2. The picture shows that the contribution has a quadratic form and its roots are located at $x_k=0$ ($x_k^{\text{measured}}-\overline{x}_k=0$) and $x_k=x_k^*$.

Table 1
Numerical example. Data reported by De Maesschalck et al. [10]

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

Table 2
Test observations

Test number	Observation	Euclidean distance to the mean
$TEST_1$	[1.000 5.350 3.125 3.245]	5.000
$TEST_2$	[11.00 5.350 3.125 3.245]	5.000
$TEST_3$	[1.000 7.000 3.125 3.245]	5.265
$TEST_4$	[11.00 7.000 3.125 3.245]	5.265
$TEST_5$	[8.000 7.000 11.00 5.000]	8.475
$TEST_6$	[2.000 8.000 8.000 7.000]	7.803
$TEST_7$	[2.100 3.100 7.900 4.900]	6.769

Eq. (13) provides a straightforward decomposition of the D-statistic into the contributions of each variable. They are obtained in the space of the original variables using the same formulation for all of them.

The proposed decomposition of the D-statistic also allows to understand the meaning of a negative variable-contribution and to estimate a bound for it. For the case shown in Fig. 2, c_k^D is negative if $0<x_k<x_k^*$. The minimum contribution value is $c_{k\min}^D$ that is located at $x_k=x_k^*/2$. If x_k is out of $0\le x_k\le x_k^*$ the D-statistic is positive and it increases with $|x_k|$. The value of variable x_k contradicts the correlation structure if $x_k<0$. On the other hand, a value of $x_k>x_k^*$ represents a large positive deviation with respect to the mean, in the direction indicated by the correlation matrix.

4. Application examples

In this section variable contributions to the D-statistic are obtained using both our proposed strategy and the method presented by Westerhuis et al. [6]. Results are compared for two case studies.

In Westerhuis approach the decomposition of the D-statistic for a new observation using a previously developed empirical model in terms of latent variables was defined as follows

$$D=\sum_{k=1}^{N}\mathbf{t}_{\text{new}}^{T}\mathbf{B}^{-1}\left[x_{\text{new},k}\mathbf{p}_k^T\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\right]^{\mathbf{T}}=\sum_{k=1}^{N}c_k^D \tag{14}$$

where $\mathbf{P}$ is the loading matrix with R retained P.C,s, $\mathbf{t}_{\text{new}}$ is the vector of scores obtained by projecting the new measurement $\mathbf{x}_{\text{new}}$ onto the model and, $\mathbf{B}$ is the covariance matrix of the score vectors corresponding to the reference population.

It should be mentioned that there exist several methods to calculate R, therefore as Valle et al. [9] mentioned, the decision to choose the R value could be very subjective. Consequently, certain amount of the available information could be lost by the projection.

4.1. Numerical example

Let us consider as reference population the one reported by De Maesschalck et al. [10] that is composed by 20 observations of four variables, as shown in Table 1. The corresponding mean vector is $\overline{\mathbf{x}}=[6\ 5.35\ 3.125\ 3.245]$. In addition, seven test observations are proposed to show how their effects on the D-statistic are interpreted by each strategy. These test observations are included in Table 2. Notice that the pairs of measurements $TEST_1/TEST_2$ and $TEST_3/TEST_4$ have the same Euclidean distance from the mean vector but some variables present deviations of different sign and magnitude.

Since the strategy developed by Westerhuis et al. [6] uses a latent variable model, two PCA of the reference data are performed considering both the covariance matrix and the correlation matrix. Figs. 3 and 4 show the total variance reconstruction and the variable variance reconstruction, respectively. In view of the results shown in Fig. 4, it is decided to perform the comparative studies using a PCA based on the correlation matrix because it provides better variable reconstructions in the plane defined by the first and second P.C.s.

In Table 3 the D-statistic values are presented for 2, 3 and 4 retained P.C.s; also the limit values for two levels of significance ($\alpha=0.05$ and $\alpha=0.01$) are provided. The statistic values that are greater than the critical ones are underlined.

The same statistic value is obtained for $TEST_1$ and $TEST_2$. It is independent of the deviation sign because the three remaining variables are at their mean values. Thus any deviation sign equally contradicts the correlation structure.

For $TEST_3$ and $TEST_4$ two variables deviate with respect to their means. The deviation of variable 2 is the same for both observations. In contrast, the deviation of variable 1 has the

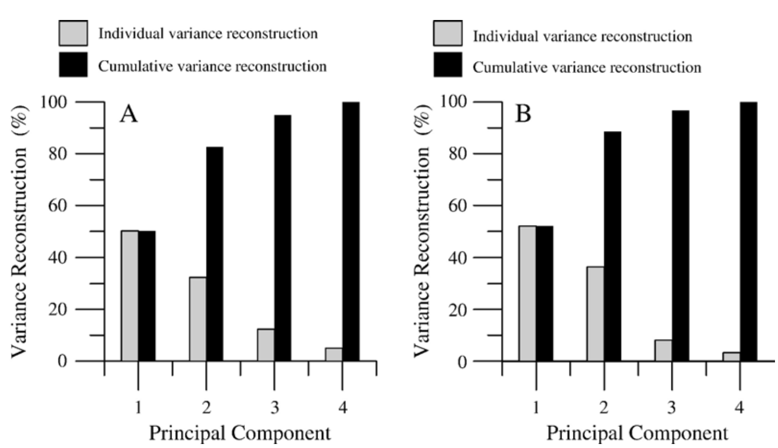

Fig. 3. Variance reconstruction using: A—the covariance matrix, B—the correlation matrix.

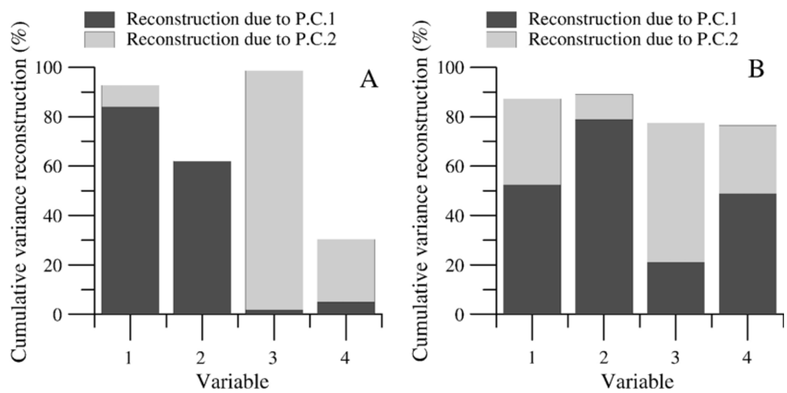

Fig. 4. Variable reconstruction: A—covariance matrix, B—correlation matrix.

same magnitude but different sign. Note that $D_{TEST3} > D_{TEST4}$ and it is also greater than the critical value when all P.C.s are taken into account, that is, when there is no lost of information regarding the correlation matrix structure.

The difference between $TEST_3$ and $TEST_4$ arises because the sign of $TEST_3$'s deviation contradicts more the correlation structure. Furthermore it should be noticed that the value of D_{TEST3} remains lower than the critical value and also lower than D_{TEST4} when the number of retained P.C.s is reduced.

For the observation $TEST_5$, D_{TEST5} is greater than the statistic critical value for $\alpha=0.05$ but not for $\alpha=0.01$. Regarding $TEST_6$, D_{TEST6} is larger than the critical value of the statistic for both levels of significance. The D-statistic value for $TEST_7$ decreases slightly when the number of retained P.C.s. diminishes. When two of them are considered, a false alarm is produced for $\alpha=0.05$.

Table 4 shows the values of the variable contributions to the statistic for each observation. They are calculated using our methodology presented in Section 3, which is identified as O.S. (original space) and, the strategy based on latent variables for 4, 3 and 2 retained P.C.s.

Table 3
Critical D values for different number of retained principal components

Observation	Retained components	D value	Critical D value ($\alpha=0.05$)	Critical D value ($\alpha=0.01$)
$TEST_1$, $TEST_2$	4	11.92	14.99	23.80
$TEST_1$, $TEST_2$	3	2.852	11.25	18.25
$TEST_1$, $TEST_2$	2	1.718	7.88	13.33
$TEST_3$	4	24.49	14.99	23.80
$TEST_3$	3	2.198	11.25	18.25
$TEST_3$	2	0.702	7.880	13.33
$TEST_4$	4	5.832	14.99	23.80
$TEST_4$	3	4.138	11.25	18.25
$TEST_4$	2	3.315	7.880	13.33
$TEST_5$	4	15.36	14.99	23.80
$TEST_5$	3	15.32	11.25	18.25
$TEST_5$	2	10.22	7.88	13.33
$TEST_6$	4	27.42	14.99	23.80
$TEST_6$	3	20.34	11.25	18.25
$TEST_6$	2	14.74	7.880	13.33
$TEST_7$	4	10.88	14.99	23.80
$TEST_7$	3	10.12	11.25	18.25
$TEST_7$	2	10.12	7.880	13.33

For observations $TEST_1$ and $TEST_2$, only variable 1 deviates from its mean value. Consequently the statistic value is equal to the contribution of that variable.

Regarding the observation $TEST_3$, $c_1^D > c_2^D$ for all the analysed cases. This indicates $\mathbf{x}_1$ deviation contradicts more the correlation structure than x_2. Although observations $TEST_3$ and $TEST_4$ are at the same Euclidean distance with respect to the mean, $TEST_4$ originates a lower violation of the correlation structure. Consequently x_1 and x_2 contributions to D_{TEST4} are lower than the corresponding values for D_{TEST3} in the original space and also when all the P.C.s are retained.

As it is shown in Table 4, the proposed strategy provides the same variable contribution values as the method by Westerhuis et al. [6] when all P.C.s are retained. This fact is demonstrated in Appendix 1. Also it can be observed that the sign of variable contributions sometimes change for different numbers of retained P.C.s.

Table 4
Variable contributions to D (c_k^D)

Observation	c_1^D	c_2^D	c_3^D	c_4^D	D	Strategy
$TEST_1$, $TEST_2$	11.92	0.000	0.000	0.000	11.92	4 P.C.s
$TEST_1$, $TEST_2$	2.852	0.000	0.000	0.000	2.852	3 P.C.s
$TEST_1$, $TEST_2$	1.718	0.000	0.000	0.000	1.718	2 P.C.s
$TEST_1$, $TEST_2$	11.92	0.000	0.000	0.000	11.92	O.S.
$TEST_3$	16.59	7.906	0.000	0.000	24.49	4 P.C.s
$TEST_3$	2.367	−0.169	0.000	0.000	2.198	3 P.C.s
$TEST_3$	1.065	−0.362	0.000	0.000	0.702	2 P.C.s
$TEST_3$	16.59	7.906	0.000	0.000	24.49	O.S.
$TEST_4$	7.256	−1.425	0.000	0.000	5.832	4 P.C.s
$TEST_4$	3.337	0.801	0.000	0.000	4.138	3 P.C.s
$TEST_4$	2.371	0.944	0.000	0.000	3.315	2 P.C.s
$TEST_4$	7.256	−1.425	0.000	0.000	5.832	O.S.
$TEST_5$	1.024	−0.233	14.97	−0.402	15.36	4 P.C.s
$TEST_5$	0.7743	0.121	15.10	−0.682	15.32	3 P.C.s
$TEST_5$	−0.187	0.477	6.917	3.016	10.22	2 P.C.s
$TEST_5$	1.024	−0.233	14.97	−0.402	15.36	O.S.
$TEST_6$	9.872	7.986	1.292	8.266	27.42	4 P.C.s
$TEST_6$	3.465	0.681	0.239	15.96	20.34	3 P.C.s
$TEST_6$	1.449	0.081	5.553	7.662	14.74	2 P.C.s
$TEST_6$	9.872	7.986	1.292	8.266	27.42	O.S.
$TEST_7$	0.582	3.290	3.905	3.105	10.88	4 P.C.s
$TEST_7$	2.626	1.261	4.242	1.996	10.13	3 P.C.s
$TEST_7$	2.657	1.252	4.156	2.056	10.12	2 P.C.s
$TEST_7$	0.582	3.290	3.905	3.105	10.88	O.S.

Table 5
Tubular reactor

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Mean	2.108	4.060	1.948	6.243	6.248
Standard deviation	0.077	0.142	0.124	0.106	0.010

	Variable 6	Variable 7	Variable 8	Variable 9	Variable 10
Mean	0.997	6.346	6.289	6.270	5.728
Standard deviation	0.027	0.012	0.009	0.009	0.232

Mean and standard deviation of the scaled data.

Furthermore it can be demonstrated that the proposed strategy reproduces the results provided by Westerhuis's technique for a lower number of retained P.C.s, when an appropriate covariance matrix is used. In this case the minimum variable-contribution concept is maintained, allowing a clearer interpretation of the sign of variable contributions to the D-statistic.

4.2. Tubular reactor example

Let us consider 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 temperature, refrigerant temperature, inlet flowrate, reactor temperature at axial positions of 10 m and 20 m, outlet reactor temperature and outlet composition of C compound, which are identified as variables 1 to 10, respectively.

Thirty seven runs are used to build the reference population. The mean and standard deviation of the scaled data are shown in Table 5, and the correlation matrix is included in Table 6. Four additional simulated runs are considered to show the performance of the strategies to identify the variables which deviations are considered in each test run, and not to isolate the source of the faults. For run (A1) an increment in the composition of component C in the feed is simulated. An increase in the outlet temperature is considered in run (A2). The third run, (A3), shows a reduction in both the refrigerant temperature and the reactor temperature at 10 m as well as a high value on the C concentration at the reactor outlet. The last run, (A4), corresponds to a reduction in the outlet temperature and an increment in the inlet concentration of component C.

Variable contributions to the D-statistic, calculated using O.S. strategy, are presented in Table 7 for each run the D-statistic is greater than the critical values ($D_{\alpha=0.05}=30.2$, $D_{\alpha=0.01}=41.9$ for $I=37$ and $N=10$), consequently an alarm arises in the space of the original variables.

A PCA of the same data gives that a total variance reconstruction of 75.4% is achieved when three P.C.s are retained, using the Cattell's criterion [1]. If only two P.C.s are retained, this value is reduced to 64%.

Table 8 contains variable contribution values calculated using the strategy based on latent variables when three P.C.s. are considered in the analysis. Only the statistic value for (A1) is greater than the critical values ($D_{\alpha=0.05}=9.404$, $D_{\alpha=0.01}=14.405$). To highlight the effects of an incorrect selection of R, an analysis considering just two retained P.C.s is included. In this case, none of the observations originate an alarm, as Table 9 indicates ($D_{\alpha=0.05}=6.903$, $D_{\alpha=0.01}=11.130$).

In Tables 7 to 9, the greatest variable contribution for each run is indicated in bold. As can be seen in Table 7, for (A1) the main contribution corresponds to the third variable (inlet C concentration), which is consistent with the actual simulated deviation The same result is obtained using the other strategy (see Tables 8 and 9).

For (A2) the decomposition in the original variable space identifies a problem in the outlet reactor temperature (variable 9) but this is not pointed out correctly in the latent variable space.

Table 6
Tubular reactor

Var	1	2	3	4	5	6	7	8	9	10
1	1.000	0.210	0.072	−0.584	−0.015	0.207	0.556	0.332	0.053	0.948
2	0.210	1.000	0.257	−0.468	−0.036	0.135	0.229	0.052	−0.105	0.376
3	0.072	0.257	1.000	−0.269	0.102	0.136	0.285	0.280	0.112	0.119
4	−0.584	−0.468	−0.269	1.000	0.241	0.011	−0.648	−0.135	0.148	−0.485
5	−0.015	−0.036	0.102	0.241	1.000	0.135	0.560	0.728	0.987	0.214
6	0.207	0.135	0.136	0.011	0.135	1.000	0.081	0.071	0.115	0.271
7	0.556	0.229	0.285	−0.648	0.560	0.081	1.000	0.703	0.648	0.614
8	0.332	0.052	0.280	−0.135	0.728	0.071	0.703	1.000	0.760	0.465
9	0.053	−0.105	0.112	0.148	0.987	0.115	0.648	0.760	1.000	0.246
10	0.948	0.376	0.119	−0.485	0.214	0.271	0.614	0.465	0.246	1.000

Correlation matrix.

Table 7
Variable contributions to D (c_k^D) calculated into the original variable space (OS)

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
A1	3.040	−0.020	**336.8**	3.770	39.85	0.540	21.59	2.770	−59.70	−3.290
A2	−356.0	−33.10	−0.100	−207.0	−1486	5.300	74.70	4.100	**1294**	805.3
A3	−523.0	35.70	1.000	−273.0	**955.8**	−0.900	504.0	16.20	−1287	680.0
A4	11.75	−111.0	2.890	34.64	−912.0	1.320	−83.80	3.120	**736.6**	358.4

The greatest variable contribution for each run is indicated in bold.

Table 8
Variable contributions to D (c_k^D) with 3 P.C.s

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
A1	0.436	−0.010	**113.1**	0.125	−0.036	0.084	−0.14	−0.180	0.019	0.204
A2	0.035	0.007	−0.010	0.062	0.081	−0.030	**0.245**	−0.150	0.012	−0.06
A3	0.133	0.098	−0.150	0.259	0.097	0.0435	**0.3343**	−0.189	0.116	0.011
A4	−0.010	0.142	−0.060	0.081	0.157	0.083	**0.250**	−0.160	0.155	0.106

The greatest variable contribution for each run is indicated in bold.

Table 9
Variable contributions to D (c_k^D) with 2 P.C.s

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
A1	−0.045	−0.001	**2.981**	0.025	0.014	−0.010	−0.045	−0.034	0.010	−0.026
A2	−0.060	−0.099	−0.007	0.275	**0.868**	−0.032	−0.062	0.412	0.702	0.001
A3	−0.035	0.007	−0.010	0.062	0.081	−0.033	**0.245**	−0.146	0.119	−0.057
A4	−0.003	0.132	−0.051	0.079	0.163	0.085	**0.249**	−0.159	0.155	0.109

The greatest variable contribution for each run is indicated in bold.

For Case (A3), Table 7 shows that the major contributions to the D-statistic are provided by the refrigerant temperature (variable 5), the outlet C concentration (variable 10) and the temperature at 10 m (variable 7). When 3 or 2 P.C.s are considered, the only important contribution is attributed to the temperature at 10 m.

The variable contributions for Case (A4) are presented in Table 7 and they indicate problems with the temperature and C concentration at the reactor outlet (variables 9 and 10). This situation is not highlighted in the results shown in Tables 8 and 9.

It should be mentioned that some of the alarms that are no fired by the D-statistic in the latent variable space could be eventually detected by the Square Prediction Error (*SPE*)-Statistic, which is used for monitoring the residual part in the PCA model. However this work is just devoted to analyze the performance of the proposed D-statistic decomposition.

5. Conclusions

In this work a strategy to decompose the D-statistic as the sum of variable contributions is proposed, that can be applied in the original variable space. The proposed decomposition provides a clear understanding of positive and negative variable-contributions and estimates a bound for the negative ones.

The estimation of each variable contribution is the same obtained by other authors in the latent variable space when all the principal components are retained. The proposed technique can be applied to monitor measured variables in the original space using only the D-statistic, without the lost of information originated by the projection into an incorrectly dimensioned P.C. space that may lead to detection faults (Type II Error), as for example in the case of $TEST_3$, and false alarms (Type I Error) as in $TEST_7$.

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. Since the method uses the inverse of this matrix, only an analysis in the latent variable space can be performed. Nevertheless, it must be kept in mind that the results are obtained based on an unreliable correlation matrix. Otherwise if the correlation matrix is well conditioned, the interpretation of monitoring results using the proposed technique is straightforward independently of the number of variables.

Acknowledgments

The authors wish to thank the financial support of CONICET (National Research Council of Argentina), ANPCyT (National Agency for the Science and Technological Promotion) and UNS (Universidad Nacional del Sur, Bahía Blanca, Argentina).

Appendix 1

Given the empirical model in terms of the whole set of latent variables

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}}, \tag{A1}$$

where $\mathbf{X}$ is the data matrix and, $\mathbf{T}$ and $\mathbf{P}$ represent the matrices of scores and loadings respectively, the contribution of a new measured variable to the D-statistic is calculated by Westerhuis et al. [6] as follows

$$c_k^D = \mathbf{t}_{\text{new}}^{\mathbf{T}}\mathbf{B}^{-1}\left[x_{\text{new},k}\mathbf{p}_k^{\mathbf{T}}\left(\mathbf{P}^{\mathbf{T}}\mathbf{P}\right)^{-1}\right]^{\mathbf{T}} \tag{A2}$$

The inverse of the covariance matrix of scores ($\mathbf{B}^{-1}$) is defined as

$$\mathbf{B}^{-1} = \left(\frac{\mathbf{T}^{\mathbf{T}}\mathbf{T}}{I-1}\right)^{-1} \tag{A3}$$

where

$$\mathbf{T} = \mathbf{X}\mathbf{P} \tag{A4}$$

and consequently,

$$\mathbf{t}_{\text{new}}^{\mathbf{T}} = \mathbf{x}_{\text{new}}^{\mathbf{T}}\mathbf{P}. \tag{A5}$$

If the empirical model is based on PCA then, $\mathbf{P}^{\mathbf{T}}=\mathbf{P}^{-1}$ because $\mathbf{P}$ is an orthonormal matrix. In consequence, Eq. (A2) can be reformulated as

$$c_k^D = \mathbf{x}_{\text{new}}^{\mathbf{T}} \mathbf{P} \left(\frac{\mathbf{P}^{\mathbf{T}} \mathbf{X}^{\mathbf{T}} \mathbf{X} \mathbf{P}}{I-1} \right)^{-1} \left[x_{\mathbf{new},k} \mathbf{p}_k^{\mathbf{T}} \right]^{\mathbf{T}} = \mathbf{x}_{\text{new}}^{\mathbf{T}} \mathbf{P} \left(\mathbf{P}^{\mathbf{T}} \mathbf{S} \mathbf{P} \right)^{-1} \mathbf{p}_k x_{\text{new},k} \tag{A6}$$

As matrices $\mathbf{P}^{\mathbf{T}}$, $\mathbf{S}$ and $\mathbf{P}$ are square and non-singular, the reversal rule for inverse products can be applied, consequently

$$c_k^D = \mathbf{x}_{\text{new}}^{\mathbf{T}} \mathbf{P} \mathbf{P}^{-1} \mathbf{S}^{-1} \left(\mathbf{P}^{\mathbf{T}} \right)^{-1} \mathbf{p}_k x_{\text{new},k} \tag{A7}$$

$$c_k^D = \mathbf{x}_{\text{new}}^{\mathbf{T}} \mathbf{S}^{-1} \mathbf{P} \mathbf{p}_k x_{\text{new},k} \tag{A8}$$

The product of $(\mathbf{P}\mathbf{p}_k)$ is a column vector with only one non-zero element for position k, thus the contribution is evaluated as follows

$$c_k^D = \left(\sum_{j=1}^{N} x_{\text{new},j} a_{jk} \right) x_{\text{new},k} = a_{kk} x_k^2 + \left(\sum_{\substack{j=1 \\ j \neq k}}^{N} a_{kj} x_j \right) x_k = a_{kk} \left(x_k^2 + \frac{\sum_{\substack{j=1 \\ j \neq k}}^{N} a_{kj} x_j}{a_{kk}} x_k \right) \tag{A9}$$

Eq. (A9) is the same obtained for the proposed strategy in the original variable space.

References

[1] J.E. Jackson, A User's Guide to Principal Components, John Wiley & Sons Inc., New York, 1991.

[2] T.C. Jeffery, J.C. Young, Electrochemical Engineering Society, in: T. Jeffery, K. Ota, J. Fenton, H. Kawamoto (Eds.), Proceedings of the Symposia of Chlor-alkali and Chlorate Production and New Mathematical and Computational Methods in Electrochemical Engineering, 16–21 May 1993, Honolulu, Hawaii, vol. 93–14, The Electrochemical Engineering Society Inc., Pennington, NJ, 1993, pp. 136–147.

[3] P. Miller, R. Swanson, C. Heckler, Appl. Math. Comput. Sci. 8 (1998) 775–792.

[4] J.F. MacGregor, C. Jaeckle, C. Kiparissides, M. Koutoudi, AIChE J. 40 (1994) 826–838.

[5] P. Nomikos, ISA Trans. 35 (1996) 259–266.

[6] J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Chemometr. Intell. Lab. Syst. 51 (2000) 95–114.

[7] R.L. Mason, N.D. Tracy, J.C. Young, J. Qual. Technol. 27 (1995) 99–108.

[8] R.L. Mason, N.D. Tracy, J.C. Young, J. Qual. Technol. 29 (1997) 396–406.

[9] S. Valle, W. Li, S.J. Qin, Ind. Eng. Chem. Res. 38 (1999) 4389–4401.

[10] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemometr. Intell. Lab. Syst. 50 (2000) 1–18.