A Nearest In Control Neighbour Based Method to Estimate Variable Contributions to the Hotelling's Statistic

Carlos R. Alvarez[a,b], Adriana Brandolín[a], Mabel C. Sánchez[a], Luis Puigjaner[b]

[a]*Planta Piloto de Ingeniería Química,BahiaBlanca 8000, Argentina*
[b]Universitat Politècnica de Catalunya, Barcelona 08029, Spain

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 (PCA) or independent component analysis (ICA). In this work we present a novel strategy based on finding the nearest "in-control" neighbour of the observation point.

Introduction

Statistical process monitoring involves three activities: detection of the out-of-control or faulty status, identification of the variable or variables that signal such condition, and diagnosis of the source cause for the abnormal behaviour. Although Hotelling's statistic is widely used to reliably detect out of control status it offers no assistance in the identification stage. A number of strategies have been proposed to assign variable-contribution values to the T^2-statistic taking into account the multivariate nature of process data.

Mason *et al.* [1,2] proposed to decompose the T^2-statistic value as a summation of J independent parts (where J 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, …, J) of the sum is the j^{th} measurement adjusted by using estimates of the mean and standard deviation of its conditional probability distribution given the (j−1) previously considered variables. Since there exists no fixed order variables, J! different but non-independent partitions can be obtained. As a possible solution for this problem, 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 contributions 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 recently presented by Alvarez *et al.* [3]. This method also provides a clear understanding of positive and negative contributions which often results from the techniques mentioned here and, estimates a bound for the negative ones.

Among the methods that work in latent variable-spaces, it was Jackson [4] who first proposed the decomposition of the T^2-statistic into a sum of principal components and perform the identification in terms of the weight of each variable in the of out-of-control component. However, in most of the industrial applications it results very difficult to associate a physical meaning to each principal component and, the variables associated with out-of-control signals cannot be determined easily. Miller *et al.* [5] and MacGregor *et al.* [6] proposed to evaluate the contributions of each process variable to the scores that are outside of their confidence limits. Nomikos [7] presented an approach to calculate the contributions of each process variable to the T^2-statistic instead of to the scores, when latent variables cannot be associated to a meaningful group of process variables. Westerhuis *et al.* [8] 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.

In all the above mentioned methods, the contribution to the T^2-statistic for each variable is estimated considering the remaining *J-1* variables fixed at their measured values. As a result of this, there is a sole "parametric curve" defining all the possible values of the T^2-statistic as function of only the analyzed variable, as it was pointed out by Alvarez *et al.* [3].

Variable contributions and contribution plots

Let's consider a chemical process in which J variables are measured and monitored over the time, and let $\mathbf{x}$ be a process observation vector containing all of those measurements for a given time instant t. The value of the Hotelling's statistic for $\mathbf{x}$ can be estimated by:

$$T^2 = (\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\overline{\mathbf{x}}), \tag{1}$$

where $\overline{\mathbf{x}}$ is the estimation 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 $\mathbf{x}$ follows a normal multivariate distribution ($\mathbf{x} = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$), then T^2 follows a $[J(I^2-1)/(I^2-IJ)]F_{J,I-J,\alpha}$ distribution, where, $F_{J,I-J,\alpha}$ is the value of the F distribution for a level of significance α, with J and $(I-J)$ degrees of freedom and I is the number of observations of the reference population.

As can be seen in equation (2), the Hotelling statistic has a squared form with the minimum at $\mathbf{x}=\overline{\mathbf{x}}$. Since matrix $\mathbf{S}$ is positive semidefinite, all the possible values for $\mathbf{x}$ will generate statistic's values that are greater than or equal to zero. For the sake of simplicity in nomenclature, the observation vectors $\mathbf{x}$ will be considered to be standardized so that $\mathbf{x} \approx N(\mathbf{0},\mathbf{R})$, where $\mathbf{R}$ is the correlation matrix.

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

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

This particular structure can be exploited in order to estimate the influence of each measurement in the final statistic's value.

Mason *et al.* [1,2] proposed to decompose the T^2-statistic value as a summation of J independent parts:

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

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

Alvarez *et al.* [3] have recently presented an alternative straightforward method to decompose the T^2-statistic as a unique sum of variable contributions as:

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

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

where $x_j^*/2$ is the x_j value that minimizes the value of T^2 given the remaining $J-1$ variables values. This decomposition of the Hotelling's 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 $0\leq x_j \leq x_j^*$. The minimum contribution value is $c_k^{\min}$ that is located at $x_j = x_j^*/2$. If x_j is out of $0\leq x_j \leq x_j^*$ the T^2-statistic is positive and it increases with $|x_j|$. The value of variable x_j contradicts the correlation structure if $x_j \leq 0$. 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 to calculate the variables' contributions when latent variables projection methods are used have bee presented as well. Nomikos [7] proposed an approach to calculate the contributions of each process variable to the Hotellings' statistic (called D when projection methods are used) instead of to the scores when latent variables cannot be associated to a meaningful group of process variables. Westerhuis *et al.* [8] presented an extended approach of the contribution plots proposed by Nomikos [7] to be applicable also to latent variable models with correlated scores. They proposed to calculate the contribution of the j^{th} variable to the inflated statistic value (c_j) as:

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

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

$$D^2 = \sum_{j=1}^{J}\mathbf{t}^\mathbf{T}\mathbf{S}_\mathbf{L}^{-1}[\hat{x}_j\mathbf{p}^\mathbf{T}]^\mathbf{T} = \sum_{n=1}^{N}c_j \ , \tag{9}$$

where $\mathbf{t}$ and $\mathbf{S_L}$ are the coordinate of $\hat{\mathbf{x}}$ in the considered latent space and variance-covariance matrix respectively. The main drawback of this technique is that it associates no interpretation to the negative values that can result from the calculations.

A Nearest In Control Neighbour Based Method (NICN)

Without assuming any gross error in measurements it still being possible to obtain an observation point whose statistic value exceeds the critical distribution value for a given significance level. This is possible if one or more variables in the observation vector does not behave as the observations in the reference population do (i.e. an anomalous event is really happening). Since one or more variables can be causing high values in the Hotelling's statistic, the question of which are the faulty variables has to be answered. One possible answer to that question could be given based in the knowledge of the nearest neighbour of the observation point that is in control. This information gives us an idea of how far from an in control allocation the faulty observation is and, which the minimum distance and direction that should be explained by an anomalous situation is. The problem of finding this nearest neighbour can be stated as an optimization problem, where the objective is to find an alternative point that minimizes a distance function to the measured point, subject to the constrain that the corresponding T^2-value is equal to or less than T_C^2.

$$\begin{aligned} &\min \quad (\mathbf{x}-\hat{\mathbf{x}})^\mathbf{T}\mathbf{\Psi}^{-1}(\mathbf{x}-\hat{\mathbf{x}}) \\ &s.t. \end{aligned} \tag{10}$$

$$\mathbf{x}^\mathbf{T}\mathbf{\Sigma}^{-1}\mathbf{x} \le T_C^2 \ , \tag{11}$$

where $\mathbf{\Psi}$ is the matrix that defines the type of distance chosen to measure the proximity to the faulty observation and, $\mathbf{\Sigma}$ is the correlation matrix estimated from the reference population.

The first order optimality conditions for the constrained optimization problem stated in equation (10) result:

$$L = (\mathbf{x} - \hat{\mathbf{x}})^{\mathbf{T}} \mathbf{\Psi}^{-1} (\mathbf{x} - \hat{\mathbf{x}}) - \lambda (\mathbf{x}^{\mathbf{T}} \mathbf{\Sigma}^{-1} \mathbf{x} - T_C^2) \tag{12}$$

$$L = \mathbf{x}^{\mathbf{T}} \mathbf{\Psi}^{-1} \mathbf{x} - \hat{\mathbf{x}}^{\mathbf{T}} \mathbf{\Psi}^{-1} \hat{\mathbf{x}} + 2 \hat{\mathbf{x}}^{\mathbf{T}} \mathbf{\Psi}^{-1} \mathbf{x} - \lambda (\mathbf{x}^{\mathbf{T}} \mathbf{\Sigma}^{-1} \mathbf{x} - T_C^2) \tag{13}$$

$$\frac{\partial L}{\partial \mathbf{x}} = 2 \mathbf{\Psi}^{-1} \mathbf{x} + 2 \lambda \mathbf{\Sigma}^{-1} \mathbf{x} - 2 \mathbf{\Psi}^{-1} \hat{\mathbf{x}} \tag{14}$$

$$\frac{\partial L}{\partial \lambda} = \mathbf{x}^{\mathbf{T}} \mathbf{\Sigma}^{-1} \mathbf{x} - T_C^2 \tag{15}$$

If we choose the distance measure to be minimized as the Mahalanobis distance between the observation and its nearest in control neighbour (i.e. $\mathbf{\Psi} = \mathbf{\Sigma}$) equations (14) and (15) can be written as:

$$\frac{\partial L}{\partial \mathbf{x}} = 2(1 + \lambda) \mathbf{\Psi}^{-1} \mathbf{x} \tag{16}$$

$$\frac{\partial L}{\partial \lambda} = \mathbf{x}^{\mathbf{T}} \mathbf{\Psi}^{-1} \mathbf{x} - T_C^2 \tag{17}$$

and the feasible points can be obtained by solving the equation system below.

$$\frac{\partial L}{\partial \mathbf{x}} = 0 \Rightarrow (1 + \lambda) \mathbf{\Psi}^{-1} \mathbf{x} = \mathbf{\Psi}^{-1} \hat{\mathbf{x}} \tag{18}$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow \mathbf{x}^{\mathbf{T}} \mathbf{\Psi}^{-1} \mathbf{x} = T_C^2 \tag{19}$$

$$(1 + \lambda) \underbrace{\mathbf{x}^{\mathbf{T}} \mathbf{\Psi}^{-1} \mathbf{x}}_{T_C^2} = \mathbf{x}^{\mathbf{T}} \mathbf{\Psi}^{-1} \hat{\mathbf{x}} \tag{20}$$

$$\lambda = \pm \left(\frac{T_C^2}{T_{Obs}^2} \right)^{1/2} \tag{21}$$

$$\mathbf{x} = \lambda \hat{\mathbf{x}} \tag{22}$$

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

The solution for the more general situation, when $\mathbf{\Psi} \neq \mathbf{\Sigma}$ is more difficult to be obtained in an analytical way. Nevertheless the computational cost of considering different distances' metrics is not much bigger.

Figure 1 shows the level curves for the two variables' case. It shows the shape and intersections between different T^2-value curves and the lines defined by the Euclidean distance (green full lines) and the Mahalanobis distance (blue dashed lines) to the out of control observation. Two different points are considered to be out of control **p₁**=[-1,2] and **p₂**=[2,0] and the reference distribution is supposed to have its mean in [0,0]. It may be noticed that the line defined for the neighbours when $\mathbf{\Psi} = \mathbf{\Sigma}$ forms a straight line. On the other hand, when the Euclidean distance is considered the intersections follows a non-linear curve.

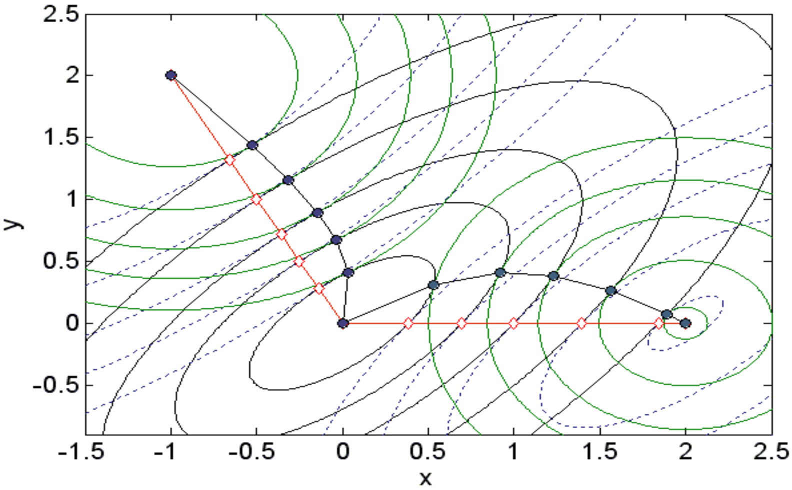

Figure 1- Different positions for the nearest in control neighbour depending on the selection of the distance metric.

The contribution of each variable to the inflated statistic value (c_j) can be estimated in several ways using this information. The easiest way is to compare the distance in which each variable should be modified to reach the nearest in control point against the average value of these distances for that observation point. Since all the variables should have been previously standardized (to be dimensionless) the comparison of the resulting movements in each direction can be used as estimates of the deviation degree of each variable, and can be used in a similar way as classically used in contribution plots. Then the directions whose changes are greater than the single threshold value (τ), are considered as suspicious variables:

$$\tau = \frac{1}{J}\sum_{j=1}^{J} c_j \tag{23}$$

Other alternatives can be used, for example, evaluating the contribution of each variable at the out of control observation as Alvarez *et al.* [3] proposed. After that, the same can be done at the resulting neighbour point. The change in each variable's contribution between these two points is an estimation of the contribution of each variable to the out of control state.

Case Studies

In this section, variable contributions to the T^2-statistic are obtained by applying the proposed strategy and compared with those estimated by using the original space strategy (OSS) (Alvarez *et al.* [3]) and the ones calculated by using equations (7) to (9). due to Westerhuis *et al.* [8]. Results for two case studies are provided.

Case Study I

Let us consider the data set formed by 20 observations of four variables presented by De Maesschalck *et al.* [9] as a reference data set. In addition, seven test observations are proposed to show how their T^2-statistic's values are interpreted by each of the three strategies (Table 1). 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.

Table 1 Test observations

	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

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. 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 same magnitude but different sign. Note that $T^2_{TEST_3} > T^2_{TEST_4}$ and it is also greater than the critical value. The difference between TEST_3 and TEST_4 arises because the sign of TEST_3's deviation contradicts more the correlation structure. For the observation TEST_5, $T^2_{\mathrm{TEST}_5}$ is greater than the statistic critical value for α=0.05 but not for α=0.01. Regarding TEST_6, $T^2_{\mathrm{TEST}_6}$ is larger than the critical value of the statistic for both levels of significance. The value of the variable contributions to the T^2 obtained by using the NICN and the OSS approaches are shown in Table 2 and Table 3 respectively. Contributions which are above their corresponding limit are underlined and in bold.

Table 2. OSS variable contributions

Observation	$c_1^{T^2}$	$c_2^{T^2}$	$c_3^{T^2}$	$c_4^{T^2}$	T^2
TEST_1 , TEST_2	<u>**11.92**</u>	0.000	0.000	0.000	11.92
TEST_3	<u>**16.59**</u>	<u>**7.906**</u>	0.000	0.000	24.49
TEST_4	<u>**7.256**</u>	-1.425	0.000	0.000	5.832
TEST_5	1.024	-0.233	<u>**14.97**</u>	-0.402	15.36
TEST_6	<u>**9.872**</u>	<u>**7.986**</u>	1.292	<u>**8.266**</u>	27.42
TEST_7	0.582	3.290	3.905	3.105	10.88

Table 3. NICN variable contributions

Observation	$c_1^{T^2}$	$c_2^{T^2}$	$c_3^{T^2}$	$c_4^{T^2}$
$TEST_1$, $TEST_2$	**0.1762** **2.0337**	0.000 0.000	0.000 0.000	0.000 0.000
$TEST_3$	**0.7847** **0.0033**	**0.3740** **0.0016**	0.000 0.000	0.000 0.000
$TEST_4$	**2.6438** **7.5539**	0.5190 1.4829	0.000 0.000	0.000 0.000
$TEST_5$	0.0001 0.0614	0.0000 0.0140	**0.0021** **0.8977**	0.0001 0.0241
$TEST_6$	**0.6693** **0.0459**	**0.5414** **0.0371**	0.0876 0.0060	**0.5604** **0.0385**
$TEST_7$	0.0176 0.1336	**0.0996** **0.7549**	**0.1182** **0.8959**	**0.0940** **0.7124**

α=0.05 (above), α=0.01 (bellow)

It can be easily noticed in the tables above that the identification capabilities of OSS and NICN are similar, only $TEST_7$ presents differences in the suspicious set of variables since no contribution exceeds its limits when OSS strategy is applied.

Since the strategy developed by Westerhuis *et al.* [8] uses a latent variable model, a PCA model of the reference data was performed considering three retained PCs, reaching a variance reconstruction over 95%. Detection capabilities of the Hotelling's statistic are strongly affected by the dimension reduction. Only $TEST_5$ and $TEST_6$ are detected as faulty observations by *D*-statistic. Consequently, only values for those will be analyzed for comparison purposes. In addition, it should be pointed out that neither $TEST_4$ nor $TEST_7$ were indicated as faulty observations for any statistic. Table 4 shows the corresponding values for D and SPE statistics. Values in bold indicate that the corresponding statistic detects a fault.

Table 4. *D* and *SPE* statistics for Case Study I when PCA is applied

	D	SPE
$TEST_1$	2.85	**1.82**
$TEST_2$	2.85	**1.82**
$TEST_3$	2.91	**4.48**
$TEST_4$	4.138	0.34
$TEST_5$	**15.31**	0.01
$TEST_6$	**20.34**	**1.42**
$TEST_7$	10.12	0.15

$D_{C,0.05}=11.25$ $SPE_{C,0.05}=0.81$
$D_{C,0.01}=18.25$ $SPE_{C,0.01}=1.48$

Table 5 shows the values for the contributions for these observations by using equations (7) to (9). For $TEST_5$, variable x_3 is pointed out as the faulty variable in concordance with the results obtained before. In contrast, when TEST$_6$ is analysed only variables x_1 and x_4 are highlighted leaving x_2 as an in-control variable.

Table 5. Variable contributions according to Westerhuis

Observation	c_1^D	c_2^D	c_3^D	c_4^D	D
TEST$_5$	0.7743	0.1206	**15.1038**	-0.6818	15.31
TEST$_6$	**3.4648**	0.6809	0.2392	**15.9576**	20.34

Case Study II

The second case study is a tubular reactor where the reaction A+B→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 of 10 m and 20 m, 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 and, four additional runs are considered to perform the same comparisons as in the previous case. For run (*R1*) an increment in the composition of component C in the feed is simulated. An increase in the outlet temperature is considered in run (*R2*). The third run, (*R3*), 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, (*R4*), corresponds to a reduction in the outlet temperature and an increment in the inlet concentration of component C.

Table 6 OSS 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	3.040	-0.020	**336.8**	3.770	39.85	0.540	21.59	2.770	-59.70	-3.290
R2	-356.0	-33.10	-0.100	-207.0	-1486	5.300	74.70	4.100	**1294**	**805.3**
R3	-523.0	35.70	1.000	-273.0	**955.8**	-0.900	**504.0**	16.20	-1287	**680.0**
R4	11.75	-111.0	2.890	34.64	-912.0	1.320	-83.80	3.120	**736.6**	**358.4**

Variable contributions to the T^2-statistic, calculated using OSS strategy, are presented in Table 6 for each run in which the D-statistic is greater than the critical values. Critical values for the OSS-based contributions were calculated as the mean value plus three standard deviations of those obtained from the reference data set.

Table 7 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	13.037	0.9767	**166.81**	13.366	43.323	0.7645	53.699	3.4070	71.829	21.995
R2	72.776	6.7550	0.0149	42.257	**303.41**	1.0724	15.178	0.8274	**264.21**	164.41
R3	117.28	7.9861	0.2316	61.078	**214.39**	0.2058	113.11	3.6351	**288.37**	152.42
R4	0.2924	2.7674	0.0722	0.8654	**22.877**	0.0328	2.0993	0.0778	**18.488**	8.9806

Table 7 shows the variables' contributions when the NICN approach is used. In this case the threshold value for contributions was set at 2 τ . Some differences appear when comparing against

Table 6. This is due to the fact that all the directions are modified when NICN approach is applied in contrast to the "fixed curve" approach given by the OSS which produce very high negative values for some contributions.

A PCA of the same data have been carried out giving a total variance reconstruction of 75.4% when three P.C.s are retained (Cattell's criterion [4] has been used to choose the number of retained PCs). When the PCA model is used to evaluate these observations, only *R1* is pointed out as an out of control observation and the variables' contributions are shown in Table 8.

Table 8 Variable contributions according to Westerhuis

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
R1	0.4363	-0.0088	**113.03**	0.1250	-0.0361	0.0840	-0.1359	-0.1765	0.0191	0.2039

As can be seen in Table 8, for (*R1*) 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 strategies.

The proposed technique has also been applied to industrial case studies, and to interpret results in latent variable spaces. More congruous results between the D-statistic variable contributions and the SPE-statistic variable contributions are obtained when the first ones are calculated using the proposed method

Conclusions

When projection techniques are used to reduce the space dimension it is possible to have either some missing or false alarms when compared with the original space results as it was shown by Alvarez *et al* [3]. This discrepancy could be due to an incorrect selection of the number of retained PCs and therefore, a strategy working in the original space can help to overcome this problem. In this work, a novel method to estimate the variable contribution to the T^2-statistic is presented. Given a measured point which T^2-value exceeds the critical value T_C^2, the contribution of each variable is determined in terms of the minimum distance between the measured point and its closer neighbour with a T^2-value equal to T^2_C. Results have shown a good performance when this technique is applied in the original variable space and they are similar to those obtained using the OSS strategy proposed by Alvarez *et al* [3]. Moreover, in other case studies it was noticed that results of NICN approach are more congruent with *SPE* contributions than those obtained using the classical approaches when PCA models are employed.

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), UNS (Universidad Nacional del Sur, Bahía Blanca, Argentina) and the European Community project MRTN-CT-2004-512233

References

1. Mason, R. L., N. D. Tracy and J. C. Young (1995), "Decomposition of T2 for multivariate control chart interpretation", *Journal of Quality Technology*, 27, 99-108.
2. Mason, R. L., N. D. Tracy and J. C. Young (1997), "A practical approach for interpreting multivariate T2 control chart signals", *Journal of Quality Technology*, 29, 396-406.
3. Alvarez, C. R., A. Brandolin and M. C. Sanchez (2007), "On the variable contributions to the D-statistic", *Chemometrics and Intelligent Laboratory Systems*, 88, 189-196.
4. Jackson, J. E. (1991), "A User's Guide to Principal Components", John Wiley & Sons, Inc., United States of America.
5. Miller, P., R. E. Swanson and C. F. Heckler (1999), "Contribution plots: The missing link in multivariative quality control", *Applied Mathematics Computer Science*, 8, 775-792.
6. Macgregor, J. F., C. Jaeckle, C. Kiparissides and M. Koutoudi (1994), "Process monitoring and diagnosis by multiblock PLS methods", *AIChE Journal*, 40, 826-838.
7. Nomikos, P. and J. F. Macgregor (1994), "Monitoring of Batch Process Using Multiway Principal Component Analysis", *AIChE Journal*, 40, 1361-1375.
8. Westerhuis, J. A., S. P. Gurden and A. K. Smilde (2000), "Generalized contribution plots in multivariate statistical process monitoring", *Chemometrics and Intelligent Laboratory Systems*, 51, 95-114.
9. De Maesschalck, R., D. Jouan-Rimbaud and D. L. Massart (2000), "The Mahalanobis distance", *Chemometrics and Intelligent Laboratory Systems*, 50, 1-18.