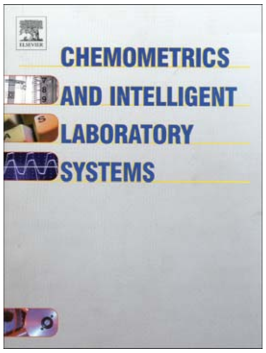
Accepted Manuscript



Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR)

Tiago J. Rato, Marco S. Reis

PII: S0169-7439(13)00059-2
DOI: doi: 10.1016/j.chemolab.2013.04.002
Reference: CHEMOM 2630

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received date: 5 October 2012
Revised date: 25 March 2013
Accepted date: 2 April 2013

Please cite this article as: Tiago J. Rato, Marco S. Reis, Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR), *Chemometrics and Intelligent Laboratory Systems* (2013), doi: 10.1016/j.chemolab.2013.04.002

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR)

Tiago J. Rato, Marco S. Reis[*]

CIEPQPF, Department of Chemical Engineering, University of Coimbra,
Rua Sílvio Lima, 3030-790, Coimbra, Portugal,

**Corresponding author: e-mail: marco@eq.uc.pt, phone: +351 239 798 700, FAX: +351 239 798 703*

Abstract

Current multivariate control charts for monitoring large scale industrial processes are typically based on latent variable models, such as principal component analysis (PCA) or its dynamic counterpart when variables present auto-correlation (DPCA). In fact, it is usually considered that, under such conditions, DPCA is capable to effectively deal with both the cross- and auto-correlated nature of data. However, it can easily be verified that the resulting monitoring statistics (T^2 and Q, also referred by SPE) still present significant auto-correlation. To handle this issue, a set of multivariate statistics based on DPCA and on the generation of decorrelated residuals were developed, that present low auto-correlation levels, and therefore are better positioned to implement SPC in a more consistent and stable way (DPCA-DR). The monitoring performance of these statistics was compared with that from other alternative methodologies for the well-known Tennessee Eastman process benchmark. From this study, we conclude that the proposed statistics had the highest detection rates on 19 out of the 21 faults, and are statistically superior to their PCA and DPCA counterparts. DPCA-DR statistics also presented lower auto-correlation, which simplifies their implementation and improves their reliability.

Keywords: Multivariate statistical process control; Principal component analysis; Dynamic Principal component analysis; Missing data imputation; Tennessee Eastman benchmark process.

1 Introduction

Current chemical process industries strive to improve the operation standards of their processes and quality levels of their products, with the central goal of reducing the variability of the main quality features around their target value [1]. Statistical Process Control (SPC) provides a toolbox for conducting such activities, where control charts have a particularly important role (so much so, that quite often the two designations, SPC and control charts, are used interchangeably). The goal of control charts is to provide a simple and objective way to monitor the process variability over time, in order to verify, at each instant, whether it remains "normal", i.e., in a state of statistical control, or whether a special cause of variation has occurred, driving it to an out of statistical control state [1, 2]. The state of statistical control is essentially characterized by process variables remaining close to their desired or average levels, affected only by common causes of variation, i.e., variation sources affecting the process all time and that are essentially unavoidable within the process normal operation conditions [1].

In these context, several SPC charts were developed for univariate processes, such as the classical Shewhart's control charts [1], CUSUM [3] and EWMA [4], and then, for multivariate (Hotelling's T^2 control chart [5], MCUSUM [6], MEWMA [7]) and megavariate (PCA-SPC, [8-10]) systems, as data and computational power becomes increasingly available. The PCA-SPC control chart, is based on a latent variables model (Principal Component Analysis, PCA), whose ability to deal with a large number of highly correlated variables is well known. It uses two complementary monitoring statistics, one of them for monitoring the variability within the PCA subspace (the Hotelling's T^2 applied to the first p latent variables, p being the process pseudo-rank) while the other follows the variability around such subspace, being a function of the projection residuals, usually referred as Q or square predicted error, SPE.

However, as for all the previous methodologies, PCA-SPC tacitly assumes that the underlying data generation process is *i.i.d.*, meaning in particular that the process mean vector is constant over time, therefore excluding any auto-correlated or non-stationary behaviour. This is currently a serious limitation, which strongly hinders the practical application of approaches based upon the *i.i.d.* assumption, due to the mass, energy and momentum inertial effects presented in most industrial systems, coupled with the high sampling rates that are currently easily achieved with modern process instrumentation and acquisition systems. To address this issue, Ku *et al.* [11] proposed an SPC procedure that extends PCA-SPC, based on a dynamical version of principal component analysis, called dynamic principal component analysis (DPCA). DPCA includes time-shifted versions of the original variables, in order to accommodate and tacitly model the dynamic behaviour of variables within the same PCA model. Unfortunately, one can easily verify that this method still leads to auto-correlated statistics, meaning that the fundamental problem raised by data autocorrelation still remains to be properly addressed.

In order to handle this issue, Rato and Reis [9] recently studied several combinations of approaches to deal with data correlation and autocorrelation, including DPCA, PLS, Time Series modelling and decorrelated residuals based on missing data imputation techniques, in a total of 22 monitoring statistics, most of them being new. From this screening study, a combination of DPCA and decorrelated residuals based on missing data imputation (DPCA-DR) stand out by their potential for dealing with data cross- and auto-correlation, and the lower levels of auto-correlation in the monitoring statistics, implying that the dynamical behavior is being properly described and incorporated in the methods' model structure. Furthermore, for the systems studied, these methods have also shown better monitoring performances when compared to their current counterparts. After such screening and characterization work, and given the good performances achieved as well as the stable monitoring behaviour of the DPCA-DR statistics, it is now both important and opportune to test in a large-scale benchmark data set, in order to consolidate the preliminary results obtained in an independently generated data set. The case study selected is the Tennessee Eastman process [12]. This case study is a widely adopted and cited benchmark in Multivariate Statistical Process Control and Fault Detection and Diagnosis, and therefore is especially suitable to test the proposed approach and to make our results comparable with those obtained from other methodologies proposed.

The rest of this article is organized as follows. In the next section, we describe the current multivariate statistics and latent variable models used in this study (PCA, and DPCA) as well as our proposed method (DPCA-DR). Next, we present and discuss the results obtained from their application to the Tennessee Eastman benchmark process, after properly defining the criteria that provide the basis for comparison. Finally, we summarize the contributions presented in this paper and present the main conclusions.

2 MethodsEquation Chapter (Next) Section 1

In this section, we briefly revise the main control chart procedures used currently for performing SPC on multivariate and megavariate processes. These methods will be used as reference against which our methodology based on DPCA-DR will be compared. We also present the new statistics based on this procedure.

2.1 Multivariate statistical process control

The natural extension of the univariate Shewhart control chart for monitoring multivariate process is the Hotelling's T^2 chart [12]. This chart assumes the process to be *i.i.d.*, following a multivariate normal distribution, and the monitoring statistic, for single observations samples is just the Mahalanobis distance between each multivariate observation and the overall reference mean. Assuming the mean and covariance matrix to be known, the monitoring statistic has the form [12-14]:

$$\chi_0^2 = \left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}-\boldsymbol{\mu}\right) \tag{1}$$

where $\mathbf{x}_{m\times 1}$ is a measurement vector, $\boldsymbol{\mu}_{m\times 1}$ is the population mean vector and $\boldsymbol{\Sigma}_{m\times m}$ is the population covariance matrix. Under multivariate normal conditions, this statistic follows a central χ^2 distribution with *m* degrees of freedom. Therefore, a multivariate χ_0^2 control chart can be constructed by plotting χ_0^2 *versus* time with an upper control limit (UCL) given by $\chi_{\alpha,n}^2$ where α is an appropriate level of significance (*e.g.* $\alpha = 0.01$) [2, 12].

When the in-control mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are unknown, they can be estimated from a sample of *n* past multivariate observations, using the usual well-known unbiased estimators of these population parameters, namely the sample mean and the sample covariance matrix [2]:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{2}$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{\mathrm{T}} \tag{3}$$

In this case, when new multivariate observations are obtained, the Hotelling's T^2 statistic is given by [2, 12, 14],

$$T^2 = \left(\mathbf{x} - \bar{\mathbf{x}}\right)^{\mathrm{T}}\mathbf{S}^{-1}\left(\mathbf{x} - \bar{\mathbf{x}}\right) \tag{4}$$

whose control chart has now the following upper control limit (UCL) [2, 14-16]:

$$UCL = \frac{m(n-1)(n+1)}{n^2 - nm}F_{\alpha,m,n-m} \tag{5}$$

where, $F_{\alpha,m,n-m}$ is the upper α percentile of the F distribution with *m* and *n* – *m* degrees of freedom. This chart is just a representative (perhaps the simplest and most well-known) of the charts that can be applied to multivariate systems of limited size (order of a dozen or less) and without strong problems of collinearity, otherwise the inversion of the covariance matrix would be highly unstable or even impossible in case of full redundancy or rank deficiency. For large scale systems, the control charts presented in the next section offer a more stable and effective solution.

2.2 Megavariate statistical process control

When the number of measured variables (*m*) becomes large (order of several dozens or higher), alternative approaches must be used, mostly due to the problems raised by the inversion of the covariance matrix in the Mahalanobis distance computation. A common solution for dealing whit this issue consists of using a latent variable modelling framework, developed for these types of processes, whose parameters can be estimated with simple and stable methods. Examples of such models are principal component analysis (PCA) [12] and partial least squares (PLS) [16], the former for problems involving a single block of variables and the latter for those where two blocks of variables need to be explicitly and simultaneously handled.

Regarding the analysis of problems with a single block of variables (the situation covered in this article), the use of PCA allows a reduction of the dimensionality of the space under monitoring, but that still preserves the essential features of the original data variability. This is achieved by transforming the original variables into a new set of uncorrelated variables, called the principal components (PCs). The first principal component (PC) of $\mathbf{x}$ is defined as the linear combination $t_1 = \mathbf{p}_1^{\mathrm{T}}\mathbf{x}$ with maximum variance subject to $\left\|\mathbf{p}_1\right\|_2 = 1$. The second PC, is the linear combination defined by $t_2 = \mathbf{p}_2^{\mathrm{T}}\mathbf{x}$ with maximum variance subject to $\left\|\mathbf{p}_2\right\|_2 = 1$, and to the condition that it must be uncorrelated with (orthogonal to) the first PC (t_1). Additional PCs are similarly defined, decomposing the entire observation matrix, $\mathbf{X}_{n\times m}$, as [12, 14]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{6}$$

where, $\mathbf{T}_{n\times p}$ is the matrix of scores, $\mathbf{P}_{m\times p}$ is the loading matrix, and $\mathbf{E}_{n\times m}$ is the residual matrix that contains the accumulated contribution of the last principal components, with small (or residual) contributions for explaining the variability exhibited by the **X** matrix. After applying the PCA decomposition to a reference data set, a Hotelling's T^2 statistic for future observations can be obtained from the first *p* PCs by,

$$T_{PCA}^2 = \sum_{i=1}^{p} \frac{t_i^2}{\lambda_i} = \mathbf{x}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Lambda}_p^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{x} \tag{7}$$

Where $\boldsymbol{\Lambda}_p$ is a diagonal matrix with the first p eigenvalues in the main diagonal and t_i is the new score for the i^{th} PC.

The upper control limit (UCL) of T_{PCA}^2 statistic is given by [11, 17]:

$$UCL = \frac{p(n-1)(n+1)}{n^2 - np} F_{\alpha,p,n-p} \tag{8}$$

where $F_{\alpha,p,n-p}$ is the upper α percentile of the F distribution with *p* and *n* – *p* degrees of freedom. Since T_{PCA}^2 only monitors the variability within the PCA subspace, spanned by the first *p* PCs, it must be complemented by a residual or lack of fit statistic, that accounts for the variation not captured by the PCA model, and monitored by T_{PCA}^2. This is achieved by computing the squared prediction error (SPE) of the residuals of a new observation ($\mathbf{e}_{m\times 1}$), also known as the Q statistic, which is defined as [11]

$$Q = \mathbf{e}^{\mathrm{T}}\mathbf{e} = \left(\mathbf{x}-\hat{\mathbf{x}}\right)^{\mathrm{T}}\left(\mathbf{x}-\hat{\mathbf{x}}\right) = \mathbf{x}^{\mathrm{T}}\left(\mathbf{I}-\mathbf{P}\mathbf{P}^{\mathrm{T}}\right)\mathbf{x} \tag{9}$$

where $\hat{\mathbf{x}}$ is the projection of a given observation onto the PCA subspace. This statistic is usually quite useful as it is sensitive to special events that cause the data to move away from the PCA subspace where normal operation data is mostly concentrated. The UCL for this statistic is given by [9],

$$UCL = \theta_1\left(\frac{c_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0\left(h_0-1\right)}{\theta_1^2}\right)^{1/h_0} \tag{10}$$

where

$$\theta_i = \sum_{j=p+1}^{n} \lambda_j^i, \quad i = 1, 2, 3 \tag{11}$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \tag{12}$$

and p is the number of retained principal components; c_α is the standard normal variable corresponding to the upper $1-\alpha$ percentile.

2.3 Megavariate statistical process control of dynamic processes

Ku *et al.* [11] showed that a linear time series relationship can be described by a conventional PCA model, through an implicit multivariate autoregressive model (VAR; processes containing moving average terms can also be approximated by finite VAR models). This is achieved by the additional inclusion of time-shifted versions of the original variables, as follows:

$$\mathbf{x}_i^{(l)} = \left[\mathbf{x}_i^{\mathrm{T}} \quad \mathbf{x}_{i-1}^{\mathrm{T}} \quad \cdots \quad \mathbf{x}_{i-l}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{13}$$

Where *i* represents an arbitrary sampling instant, *l* is the number of lags (or time-shifts) to be considered and $\mathbf{x}_i^{(l)}$ is the resulting augmented vector of variables for the instant *i*. The augmented matrix is obtained by the straightforward superposition of these lines of augmented observations, $\mathbf{x}_i^{(l)}$. Then, PCA is applied over such augmented matrix, providing a description of not only the (static) cross-correlations among variables but their auto-correlations and lagged cross-correlations, due to the additional incorporation of time-shifted variables. Furthermore, by properly choosing the number of lags to include, l, both the static and dynamic relations should appear in the noise subspace composed corresponding to the PCs with small variance [11].

2.4 Megavariate statistical process control of dynamic processes: the DPCA-DR approach

The introduction of time-shifted variables on DPCA has the purpose of describing the autocorrelation and lagged cross-correlations present in data, besides the cross-correlation features. However, looking to the behaviour of the resulting DPCA T^2 and Q statistics, one can verify that they, in fact, still present autocorrelation, and the underlying problem of removing it from the monitoring statistics, so that they can be handled with simple charts based on the *i.i.d.* assumption, was not properly solved yet. In order to mitigate this issue, Rato and Reis [9] have recently proposed a new methodology that combines a DPCA model and decorrelated residuals obtained from a conditional data imputation technique, in order to obtain better time-decorrelated statistics, in a simple way, within the same modelling approach, without the need to resource on further time-series modelling in order to compensate for the remaining dynamic patterns of the statistics. The underlying reasoning is the following. In DPCA, a matrix with current and past measurements is build. At each new incoming observation, *i*, a new observed score and projection can be computed. If one assumes that the current multivariate observation vector is missing, the associated values for the current scores and projections can still be estimated from past data using a conditional missing data imputation technique for PCA (in this case, DPCA). This essentially amounts to perform a one-step-ahead prediction of the scores and observations, obtained with an implicit latent variable VAR model estimated by DPCA. We have verified that the residuals obtained from the differences between the observed and estimated scores and projections are almost serially decorrelated, meaning that such residuals are ready to be monitored by simple control charting procedures. The conditional data imputation method chosen was the conditional mean replacement [18, 19]. In this method, a measurement vector with missing data is rearranged, without loss of generality, as follows,

$$\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} \mathbf{x}^{\#\mathrm{T}} & \mathbf{x}^{*\mathrm{T}} \end{bmatrix} \tag{14}$$

where $\mathbf{x}^{\#}$ denotes the missing measurements and $\mathbf{x}^{*}$ the observed ones. Correspondingly, the $\mathbf{P}$ matrix is also partitioned as $\mathbf{P}^{\mathrm{T}} = \begin{bmatrix} \mathbf{P}^{\#\mathrm{T}} & \mathbf{P}^{*\mathrm{T}} \end{bmatrix}$. The missing measurements ($\mathbf{x}^{\#}$) can be estimated by application of the Expectation-Maximization (EM) algorithm, where at each iteration, their values are replaced by the expected ones from the conditional normal distribution given the known data and the current estimate of the mean and covariance matrix (expectation stage), that is [18],

$$\hat{\mathbf{x}}^{\#} = E\left(\mathbf{x}^{\#} \middle| \mathbf{x}^{*}, \bar{\mathbf{x}}, \mathbf{S}\right) \tag{15}$$

which will then be used to update the model (maximization stage), and so forth, until a convergence criteria regarding the change on the successive solutions, is met. In our case, we assume that a PCA model is already available (the DPCA model built from reference data), and therefore, only the expectation step of the EM algorithm is required, in order to calculate the estimates for the missing measurements. Substituting $\mathbf{P}$ into the expression for $\mathbf{S}$ results in [18],

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^{\#}\mathbf{\Lambda}\mathbf{P}^{\#\mathrm{T}} & \mathbf{P}^{\#}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}} \\ \mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{\#\mathrm{T}} & \mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}} \end{bmatrix} \tag{16}$$

Using this expression for $\mathbf{S}$, the conditional expectation of the missing measurements is simply given by [18]

$$\hat{\mathbf{x}}^{\#} = \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{x}^{*} = \mathbf{P}^{\#}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\left(\mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{x}^{*} \tag{17}$$

The estimated measurements can then be used in the score calculations along with the observed data, as if no measurements were missing. For PCA this leads to [18]

$$\begin{aligned}\hat{\mathbf{t}} = \mathbf{P}_{1:p}^{\mathrm{T}}\begin{bmatrix}\hat{\mathbf{x}}^{\#}\\ \mathbf{x}^{*}\end{bmatrix} &= \mathbf{P}_{1:p}^{\mathrm{T}}\begin{bmatrix}\mathbf{P}^{\#}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\left(\mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{x}^{*}\\ \left(\mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\right)\left(\mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{x}^{*}\end{bmatrix}\\ &= \mathbf{P}_{1:p}^{\mathrm{T}}\begin{bmatrix}\mathbf{P}^{\#}\\ \mathbf{P}^{*}\end{bmatrix}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\left(\mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{x}^{*}\\ &= \mathbf{P}_{1:p}^{\mathrm{T}}\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\left(\mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{x}^{*}\\ &= \left[\mathbf{I}\quad\mathbf{0}\right]\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\left(\mathbf{P}^{*}\mathbf{\Lambda}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{x}^{*}\end{aligned}\tag{18}$$

where $\mathbf{P}_{1:p}$ is the matrix of the first p eigenvectors, $\mathbf{I}$ is an ($p \times p$) identity matrix and $\mathbf{0}$ is an ($p \times (m - p)$) matrix of zeros.

The same approach can be applied to DPCA for generating decorrelated residuals. In this case, we consider that the current variables, $\mathbf{x}_i^{\mathrm{T}}$ in Equation (13) are unknown. Therefore, the application of this methodology will give us an estimate of the scores that best agree with the last l known measurements. Given such estimated scores, we have defined the following Hotelling's T^2 statistic:

$$T_{PREV}^{2} = \left(\mathbf{t}\text{ - }\hat{\mathbf{t}}\right)^{\mathrm{T}}\mathbf{S}_{\mathbf{t}\text{-}\hat{\mathbf{t}}}^{-1}\left(\mathbf{t}\text{ - }\hat{\mathbf{t}}\right)\tag{19}$$

where $\mathbf{S}_{\mathbf{t}\text{-}\hat{\mathbf{t}}}$ is the sample covariance matrix of the difference between the observed and estimated scores, ($\mathbf{t}-\hat{\mathbf{t}}$), that monitors the DPCA reference subspace. Likewise, a monitoring statistic for the residual subspace can be defined as:

$$T_{RES}^{2} = \mathbf{r}^{\mathrm{T}}\mathbf{S}_{\mathbf{r}}^{-1}\mathbf{r} = \left(\mathbf{x}-\mathbf{P}\hat{\mathbf{t}}\right)^{\mathrm{T}}\mathbf{S}_{\mathbf{r}}^{-1}\left(\mathbf{x}-\mathbf{P}\hat{\mathbf{t}}\right)\tag{20}$$

where $\mathbf{S}_{\mathbf{r}}$ is the sample covariance matrix of the residuals in the reconstructed data, obtained with the estimated scores ($\mathbf{r} = \mathbf{x} - \mathbf{P}\hat{\mathbf{t}}$). These two statistics present low levels of autocorrelation and very interesting detection performances, as will be illustrated in the following section for the Tennessee Eastman case study.

3 A comparison study based on the Tennessee Eastman benchmark process

In order to test and compare the monitoring features and performance of the proposed methodology, we have selected an application scenario which has been widely used in process monitoring and fault detection studies: the Tennessee Eastman benchmark process. This case study not only provides a challenging testing environment for the specific comparison study carried out in this work, but also enables and simplifies the extension of the comparison scope to other methods tested in the same system, such as [11, 20-23]. A model of this process was developed by Downs and Vogel [10], consisting of five major transformation units, which are a reactor, a condenser, a compressor, a separator, and a stripper, as shown in Figure 1. From this model, 41 measurements (XMEAS) are generated along with 12 manipulated (XMV) variables. A total of 21 different process upsets are simulated for testing the detection ability of the monitoring methods, as presented in Table 1 [20, 24]. In the current study we have conducted our analysis with the data set used by Russell *et al*. [25] (available at http://web.mit.edu/braatzgroup), where the Tennessee Eastman process is controlled with the approach suggested by Lyman and Georgakis [24]. Each data set contains 960 observations collected at a sample interval of 3 min and the faults were introduced 8 hours after the simulations start. All the manipulated and measurement variables, except the agitation speed of the reactor’s stirrer (which is always constant), were collected, giving a total of 52 variables.

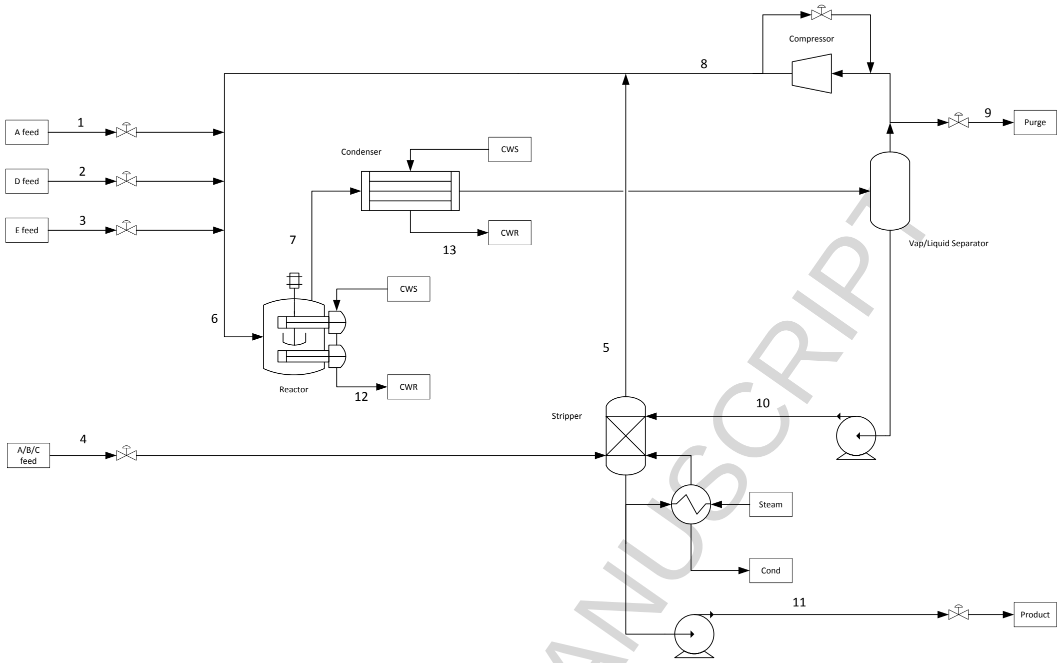


Figure 1 The Tennessee Eastman process flow sheet.

Table 1 Process faults for the Tennessee Eastman process simulator.

Variable	Description	Type
IDV(1)	A/C feed ratio, B composition constant(Stream 4)	Step
IDV(2)	B composition, A/C ratio constant (Stream 4)	Step
IDV(3)	D feed temperature (Stream 2)	Step
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss (Stream 1)	Step
IDV(7)	C header pressure loss - reduced availability (Stream 4)	Step
IDV(8)	A, B, C feed composition (Stream 4)	Random variation
IDV(9)	D feed temperature (Stream 2)	Random variation
IDV(10)	C feed temperature (Stream 4)	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water inlet temperature	Random variation
IDV(13)	Reaction kinetics	Slow drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking

IDV(16)	Unknown	
IDV(17)	Unknown	
IDV(18)	Unknown	
IDV(19)	Unknown	
IDV(20)	Unknown	
IDV(21)	The valve for Stream 4 was fixed at the steady state position	Constant position

A data set with no faults, representing normal operation conditions was used to estimate the reference PCA, DPCA and DPCA-DR models. The number of principal components for PCA and DPCA was determined by parallel analysis and the number of lags was selected by the approach proposed by Ku *et. al* [11]. Using these methods we constructed a PCA model with 17 PCs and a DPCA model with 3 lags and 29 PCs. These results are in good accordance with those obtained by Russell *et al.* [20]. For selecting the number of lags for the DPCA-DR model we have used the algorithm proposed by Rato and Reis [26]. This algorithm is based on a succession of singular value decomposition problems, and subsequent analyses of an auxiliary function from which the lags for each variable can be set. Table A.1 summarizes the number of lags considered for each variable obtained with this methodology, which led to a model with 69 PCs.

We would like to point out that the direct use of the theoretical significance levels for establishing the statistical control limits for the various methods may lead to widely different observed false alarm rates, which distorts any comparison study on the methods detection performances. This undesirable effect can be removed by manipulating the theoretical significance level of the control limits in such a way that the effectively observed performance for all methods under normal operations conditions (i.e., their false alarm rate), becomes equal. Only in such condition can all the methods be properly compared with future test data. Therefore, the UCL for the various methods were set to a false alarm rate of 1% under normal operation conditions, by trial and error, on a second data set with no faults. With this preliminary but important procedure concluded, the fault detection rates for all the methods regarding each fault were finally determined. A summary of the results obtained is presented in Table 2.

Table 2 Fault detection rates for the various methods under study, regarding each faulty scenario (a description of each fault can be found in Table 1). The top scores are signalled in boldface format.

Fault	PCA		DPCA		DPCA-DR	
	T^2	Q	T^2	Q	T^2_{PREV}	T^2_{RES}
1	0.991	0.995	0.990	0.994	0.996	**0.998**
2	**0.985**	0.984	0.984	0.981	**0.985**	0.983
3	**0.036**	0.006	0.035	0.010	0.021	0.016
4	0.218	0.980	0.165	**0.999**	0.998	**0.999**
5	0.257	0.217	0.293	0.228	**0.999**	**0.999**
6	0.989	**0.999**	0.989	**0.999**	**0.999**	**0.999**
7	**0.999**	**0.999**	0.986	**0.999**	**0.999**	**0.999**
8	0.974	0.968	0.973	0.974	**0.985**	0.981
9	**0.034**	0.010	0.030	0.002	0.020	0.010
10	0.367	0.154	0.439	0.172	**0.956**	0.933
11	0.414	0.638	0.340	0.829	**0.965**	0.865
12	0.985	0.925	0.990	0.964	**0.998**	**0.998**
13	0.943	0.950	0.943	0.950	**0.958**	0.956
14	0.988	**0.999**	0.990	**0.999**	0.998	**0.999**
15	0.035	0.007	0.059	0.009	**0.385**	0.047
16	0.174	0.137	0.217	0.145	**0.976**	0.945
17	0.787	0.905	0.790	0.953	**0.976**	0.975
18	0.893	0.901	0.890	0.898	**0.905**	0.900
19	0.115	0.059	0.046	0.298	**0.971**	0.843
20	0.340	0.423	0.408	0.493	0.908	**0.916**
21	0.362	0.414	0.429	0.409	0.539	**0.577**

From the analysis of Table 2, it is possible to verify that the DPCA-DR monitoring statistics tend to present highest fault detection rates. In fact, T^2_{PREV} was the best statistic in 14 out of 21 faults, and T^2_{RES} in 9 of them. Globally, they were capable to detect 19 of 21 faults, failing only in the detection of faults number 3 and 9, where all methods also present problems. Fault number 15 is another example of a fault difficult to detect, but where the statistic T^2_{PREV} achieved the best score. The lower capability for detecting these three specific faults was expected, as other methods reported in the literature (*e.g.* PCA, DPCA and CVA) also fail to detect them [20].

In order to better illustrate the monitoring behaviour of the methods under analysis, we present in Figures 2 and 3 the control charts for some of the process faults. From these representations it is possible to clearly observe that only the DPCA-DR statistics present a consistent out of control state in both statistics, simultaneously (T^2_{PREV} and T^2_{RES}, see

Figure 2). This is a relevant issue, since the PCA and DPCA statistics may lead to the erroneous conclusion that the process has returned to their normal operation conditions and it is no longer under the effect of a fault. In the case of Fault 10 (Figure 3) only the DPCA-DR statistics signals an out of control state during the total duration of the fault, while the PCA and DPCA statistics only became out of control when the data exceeds their normal values.

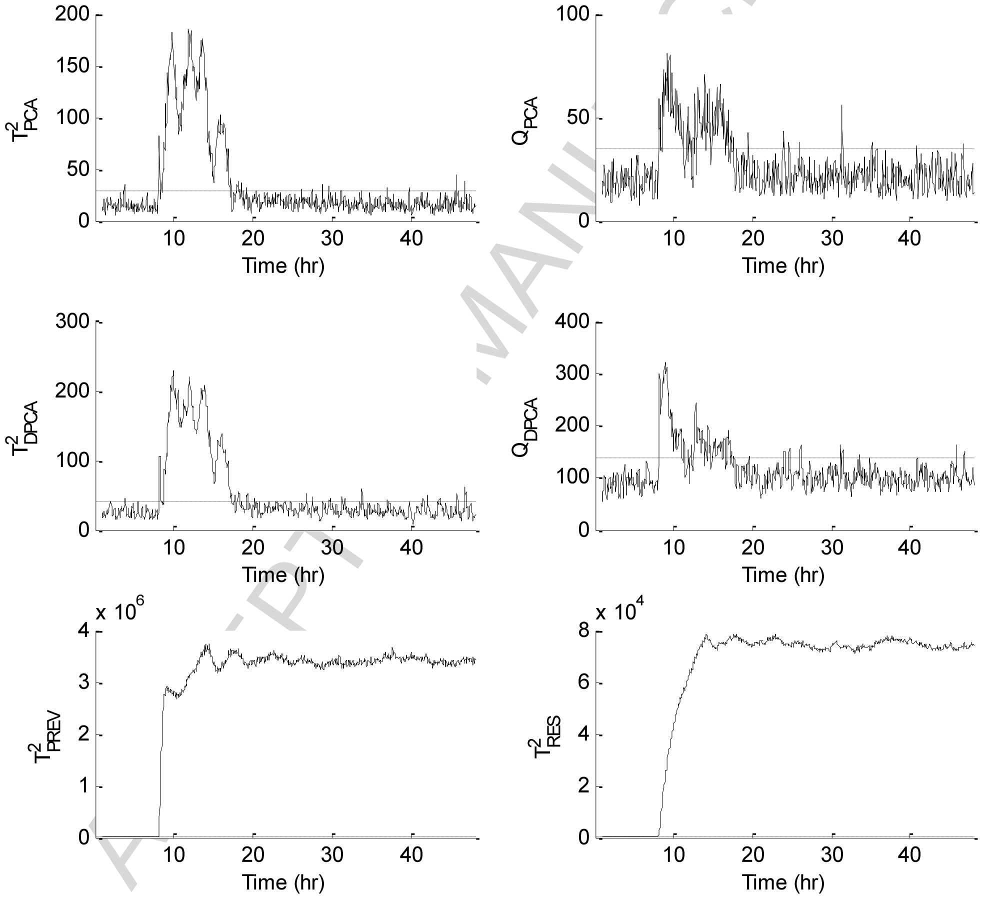


Figure 2 The multivariate statistics under test for Fault 5: PCA statistics (first or top row), DPCA statistics (second or middle row) and DPCA-DR statistics (third or bottom row).

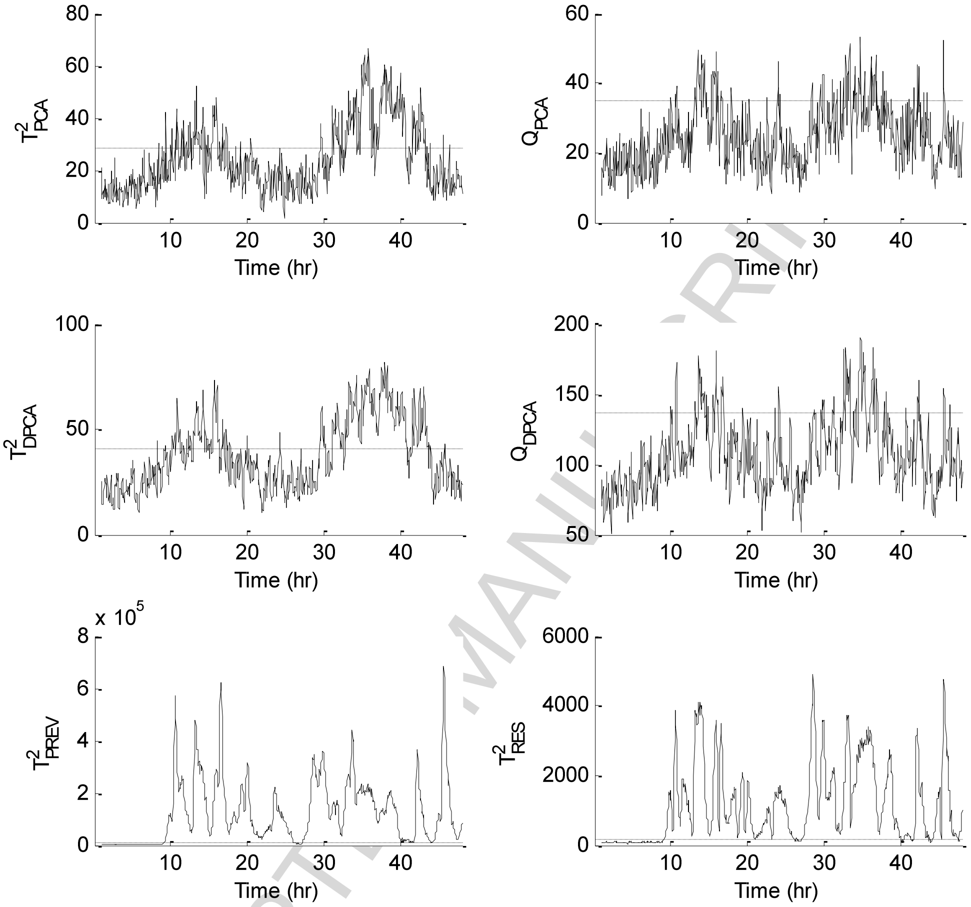


Figure 3 The multivariate statistics under test for Fault 10: PCA statistics (first or top row), DPCA statistics (second or middle row) and DPCA-DR statistics (third or bottom row).

In order to confirm the overall superiority of the DPCA-DR statistics in this case study, we have conducted paired t-tests between all the statistics presented in Table 3 (as they were implemented over the same data sets, they are paired by design in the comparison study). The test statistic is given by $t_0 = \bar{D}\Big/\left(s_D\big/\sqrt{n}\right)$, where $\bar{D}$ is the sample average of the differences between two methods under analysis in the *n* different testing conditions, D_1, D_2, … , D_n, and s_D is the sample standard deviation of these differences [27]. From this analysis it can be concluded that, with a 5% significance level, the DPCA-DR statistics are indeed significantly better than all the PCA and DPCA statistics.

Table 3 *p*-values for the paired t-test involving the detection rates obtained with method A (see first column) and method B (see first line), on all simulated faults, along with the signal of the test statistic, *i.e.* sign(t_0). For instance, a plus (+) signal, indicates that method A leads to higher detections rates, on average, when compared to method B. Values in bold indicate *p*-values lower than 0.05 (i.e., statistically significant differences at this level).

B A	PCA		DPCA		DPCA-DR	
	T^2	Q	T^2	Q	T^2_{PREV}	T^2_{RES}
PCA						
T^2		0.388 (-)	0.414 (-)	0.152 (-)	**0.002** **(-)**	**0.003** **(-)**
Q	0.388 (+)		0.540 (+)	**0.046** **(-)**	**0.004** **(-)**	**0.008** **(-)**
DPCA						
T^2	0.414 (+)	0.540 (-)		0.257 (-)	**0.002** **(-)**	**0.004** **(-)**
Q	0.152 (+)	**0.046** **(+)**	0.257 (+)		**0.007** **(-)**	**0.013** **(-)**
DPCA-DR						
T^2_{PREV}	**0.002** **(+)**	**0.004** **(+)**	**0.002** **(+)**	**0.007** **(+)**		0.115 (+)
T^2_{RES}	**0.003** **(+)**	**0.008** **(+)**	**0.004** **(+)**	**0.013** **(+)**	0.115 (-)	

Another advantage of the DPCA-DR method is the lower autocorrelation levels of its statistics, where much of its success may lie, as this characteristic makes the DPCA-DR statistics more reliable and consistent with the type of control charts used to monitor them (Figure 4).

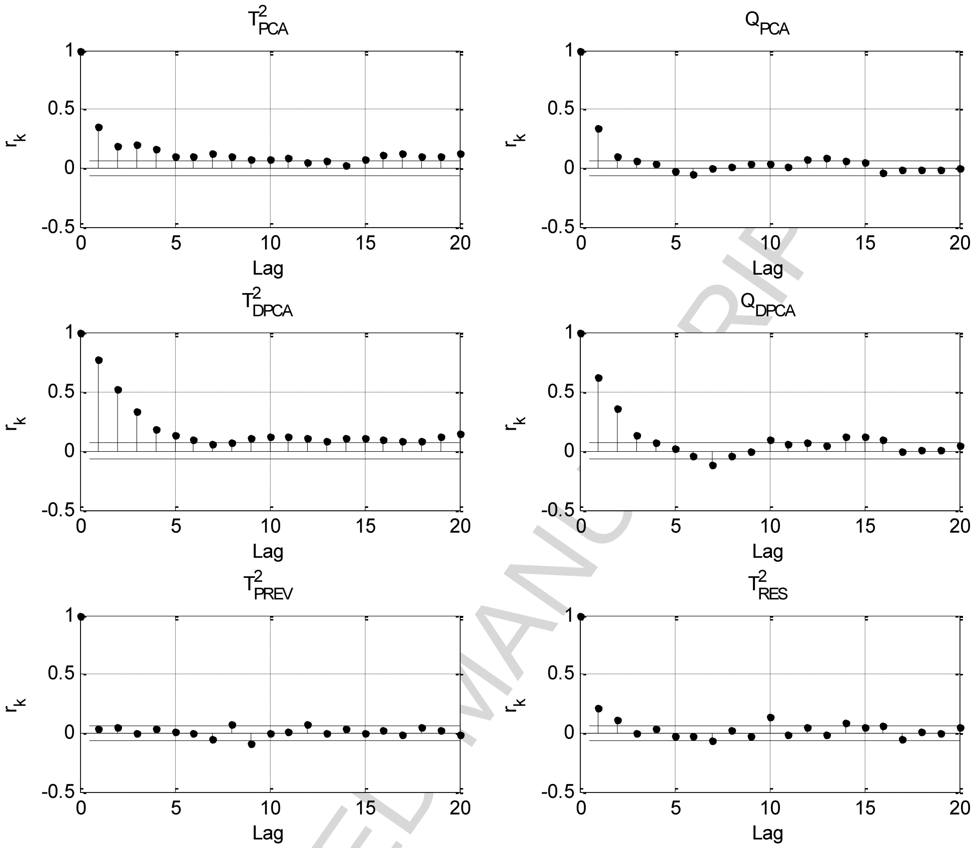


Figure 4 Auto-correlation plots for the monitoring statistics when the process is under normal operation conditions (data set with no faults). The proposed DPCA-DR statistics present the lowest levels of correlation among all the studied ones. PCA statistics - first or top row -, DPCA statistics - second or middle row -, DPCA-DR statistics - third or bottom row -.

4 Conclusions

In this paper we have presented a methodology for conducting large-scale process monitoring of dynamical systems, called DPCA-DR, and compared its performance against other well-known methodologies used in the same application context, namely PCA and DPCA. The comparison study was conducted using the Tennessee Eastman benchmark process and all faults considered in its design. From the analysis of the results obtained, we can conclude that the DPCA-DR statistics were superior, in a statistically significant sense, to the other ones, achieving the highest detection scores in 19 out of the 21 faults.

On the other hand, the DPCA-DR statistics also presented the lowest auto-correlation levels and were able to sustain the out-of-control signals during the whole faults duration, while PCA and DPCA statistics often return to their in-control regions leading to a false sense of normality. Consequently, the DPCA-DR statistics seems to be more effective, reliable and consistent regarding their counterparts tested in this study, features that make them a viable alternative to current monitoring statistics.

Acknowledgements

Tiago J. Rato acknowledges the Portuguese Foundation for Science and Technology for his PhD grant (grant SFRH/BD/65794/2009). Marco S. Reis also acknowledges financial support through project PTDC/EQU-ESI/108374/2008 co-financed by the Portuguese FCT and European Union’s FEDER through “Eixo I do Programa Operacional Factores de Competitividade (POFC)” of QREN (with ref. FCOMP-01-0124-FEDER-010397).

5 References

[1] W.A. Shewhart, Economic Control of Quality of Manufactured Product, D. Van Nostrand Company, Inc., New York, 1931.

[2] D.C. Montgomery, Introduction to Statistical Quality Control, fifth ed., Wiley, 2005.

[3] E.S. Page, Continuous Inspection Schemes, Biometrics, 41 (1954) 100-115.

[4] S.W. Roberts, Control Charts Tests Based on Geometric Moving Averages, Technometrics, 1 (1959) 239-250.

[5] H. Hotelling, The Generalization of Student's Ratio, Ann. Math. Stat., 2 (1931) 360-378.

[6] R.B. Crosier, Multivariate Generalizations of Cumulative Sum Quality-Control Schemes, Technometrics, 30 (1988) 291-303.

[7] C.A. Lowry, W.H. Woodall, C.W. Champ, S.E. Rigdon, A Multivariate Exponentially Weighted Moving Average Control Chart, Technometrics, 34 (1992) 46-53.

[8] J.E. Jackson, G.S. Mudholkar, Control Procedures for Residuals Associated With Principal Component Analysis, Technometrics, 21 (1979) 341-349.

[9] T.J. Rato, M.S. Reis, Statistical Process Control of Multivariate Systems with Autocorrelation, Submitted (2012).

[10] J.J. Downs, E.F. Vogel, A plant-wide industrial process control problem, Comput. Chem. Eng., 17 (1993) 245-255.

[11] W. Ku, R.H. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, Chemom. Intell. Lab. Syst., 30 (1995) 179-196.

[12] T. Kourti, J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemom. Intell. Lab. Syst., 28 (1995) 3-21.

[13] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, 3 ed., Wiley, New Jersey, 2003.

[14] N.D. Tracy, J.C. Young, R.L. Mason, Multivariate Control Charts for Individual Observations, J. Qual. Technol., 24 (1992) 88-95.

[15] J.E. Jackson, Quality Control Methods for Several Related Variables, Technometrics, 1 (1959) 359-377.

[16] J.V. Kresta, J.F. MacGregor, T.E. Marlin, Multivariate Statistical Monitoring of Process Operating Performance, Can. J. Chem. Eng., 69 (1991) 35-47.

[17] J.F. MacGregor, T. Kourti, Statistical process control of multivariate processes, Control Eng. Pract., 3 (1995) 403-414.

[18] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Missing data methods in PCA and PLS: Score calculations with incomplete observations, Chemom. Intell. Lab. Syst., 35 (1996) 45-65.

[19] F. Arteaga, A. Ferrer, Dealing with missing data in MSPC: several methods, different interpretations, some examples, J. Chemom., 16 (2002) 408-418.

[20] E.L. Russell, L.H. Chiang, R.D. Braatz, Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, Chemometrics and Intelligent Laboratory Systems, 51 (2000) 81-93.

[21] L.H. Chiang, R.D. Braatz, Process monitoring using causal map and multivariate statistics: fault detection and identification, Chemom. Intell. Lab. Syst., 65 (2003) 159-178.

[22] P.P. Odiowei, Y. Cao, State-space independent component analysis for nonlinear dynamic process monitoring, Chemom. Intell. Lab. Syst., 103 (2010) 59–65.

[23] S. Yin, S.X. Ding, A. Haghani, H. Hao, P. Zhang, A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process, Journal of Process Control, 22 (2012) 1567-1581.

[24] P.R. Lyman, C. Georgakis, Plant-wide control of the Tennessee Eastman problem, Comput. Chem. Eng., 19 (1995) 321-331.

[25] E.L. Russell, L.H. Chiang, R.D. Braatz, Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes, Springer, 2000.

[26] T.J. Rato, M.S. Reis, A New Methodology for Defining the Structure of Dynamic Principal Component Analysis Models and its Impact in Process Monitoring and Prediction, Submitted (2012).

[27] D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, John Wiley & Sons, Inc., 2003.

Appendix A

Table A.1 Number of lags for each variable obtained with Rato and Reis [26] lag selection method.

XMEAS(1)	17	XMEAS(14)	4	XMEAS(27)	17	XMEAS(40)	12
XMEAS(2)	17	XMEAS(15)	17	XMEAS(28)	13	XMEAS(41)	17
XMEAS(3)	8	XMEAS(16)	12	XMEAS(29)	3	XMV(1)	17
XMEAS(4)	17	XMEAS(17)	17	XMEAS(30)	17	XMV(2)	17
XMEAS(5)	17	XMEAS(18)	17	XMEAS(31)	17	XMV(3)	17
XMEAS(6)	16	XMEAS(19)	17	XMEAS(32)	8	XMV(4)	17
XMEAS(7)	17	XMEAS(20)	17	XMEAS(33)	8	XMV(5)	15
XMEAS(8)	15	XMEAS(21)	17	XMEAS(34)	17	XMV(6)	16
XMEAS(9)	17	XMEAS(22)	17	XMEAS(35)	17	XMV(7)	17
XMEAS(10)	17	XMEAS(23)	17	XMEAS(36)	17	XMV(8)	17
XMEAS(11)	16	XMEAS(24)	17	XMEAS(37)	17	XMV(9)	16
XMEAS(12)	17	XMEAS(25)	17	XMEAS(38)	17	XMV(10)	17
XMEAS(13)	17	XMEAS(26)	17	XMEAS(39)	4	XMV(11)	17

Research Highlights

- The recently proposed monitoring statistics based on Dynamic Principal Component Analysis and Missing Data imputation methods (DPCA-MD) are introduced and described.
- The monitoring performance of these statistics was compared with those from other alternative methodologies, namely PCA and DPCA.
- The system used in the comparison study is the well-known Tennessee Eastman process benchmark.
- The results obtained demonstrate the potential of the proposed monitoring statistics as valid alternatives to the current ones, has they are quite simple to implement computationally and lead to significantly better results.