I&EC research
Industrial & Engineering Chemistry Research

Article

pubs.acs.org/IECR

Advantage of Using Decorrelated Residuals in Dynamic Principal Component Analysis for Monitoring Large-Scale Systems

Tiago J. Rato and Marco S. Reis*

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

ABSTRACT: A new methodology is proposed for monitoring multi- and megavariate systems whose variables present significant levels of autocorrelation. The new monitoring statistics are derived after the preliminary generation of decorrelated residuals in a dynamic principal component analysis (DPCA) model. The proposed methodology leads to monitoring statistics with low levels of serial dependency, a feature that is not shared by the original DPCA formulation and that seriously hindered its dissemination in practice, leading to the use of other, more complex, monitoring approaches. The performance of the proposed method is compared with those of a variety of current monitoring methodologies for large-scale systems, under different dynamical scenarios and for different types of process upsets and fault magnitudes. The results obtained clearly indicate that the statistics based on decorrelated residuals from DPCA (DPCA-DR) consistently present superior performances regarding detection ability and decorrelation power and are also robust and efficient to compute.

1. INTRODUCTION

Since its introduction in the early 1930s by Shewhart, statistical process control (SPC) has been employed in industry as a useful and valuable tool for properly handling the natural or common causes of variability in daily operations and for quickly detecting process anomalies requiring intervention (special or assignable causes of variation).[1] With the continually increasing demand for improved products and processes over the years, new SPC procedures have been developed to meet the needs for higher detection sensitivities and to cope with new data structures that become available (spectra, images, hyphenated-instrument measurements, etc.), as well as with the more complex nature of industrial processes and systems. Regarding continuous production systems (the scope of this article), this trend has resulted, for instance, in the successive development of more sensitive and robust univariate methodologies, such as EWMA (exponentially weighted moving average)[2] or CUSUM (cumulative sum);[3] the extension of the Shewhart chart to multivariate full-rank systems, through multivariate statistical process control (MSPC) based on the Hotelling T^2 control chart,[4] MEWMA (multivariate exponentially weighted moving average),[5] and MCUSUM (multivariate cumulative sum);[5] and then to rank-deficient or large-scale systems, implementing MSPC (multivariate statistical process control) based on latent-variable methods, such as principal component analysis (PCA)[6,7] or partial least-squares (PLS, also known as projection onto latent structures).[8−11]

All of the mentioned SPC methodologies rely on the conventional iid (independent and identically distributed) assumption for process measurements and, therefore, are of limited application scope in most current industrial processes, where the intrinsic process dynamics (driven by mass, energy, and momentum conservation laws, as well as by the action of control systems under external perturbations) associated with high sampling rates (very easy to achieve with the current state of instrumentation technology) very frequently lead to strong violations of this hypothesis. To address this limitation, three classes of approaches have been developed: (i) control limit adjustment, (ii) embedded dynamical modeling (namely, but not only, for the generation of residuals), and (iii) data transformation techniques.

The first class of approaches consists of adjusting the control limits so that the control charts preserve the specified false alarm rate in the dynamical process under monitoring, as they would in the case of iid systems. Such corrections can be derived analytically for some simple dynamic systems and employed if the data time series is consistent with such descriptions (or dynamic models). The most common model employed is the AR (autoregressive) process (first or second order), and corrections have been derived for the common types of control charts: Shewhart,[12] CUSUM,[13] and EWMA.[14] This class of approaches is very limited by the small number of cases studied, consisting only of the simpler representatives of univariate dynamical processes.

In the second class of approaches, a dynamical model is first derived for the process under analysis using normal operation data. This model is then employed to predict the current value of the process (say, at time k), using data acquired until the previous sampling time (time $k-1$). The difference between the estimated and measured values is finally computed at time k, and if the estimated model is appropriate, it should be approximately normal (under Gaussian assumptions) and independently distributed, with zero mean and constant variance. Several explicit model structures can be used for such purposes, with the aim of removing the autocorrelation trend, such as time-series models, namely, ARIMA (autoregressive integrated moving average);[15,16] state-space models;[17−19] and dynamic latent-variable models, such as DPCA[20] or DPLS.[21] Despite requiring an additional stage of modeling, for which the necessary skills must be present in practice in the

Received: December 19, 2012
Revised: August 24, 2013
Accepted: August 26, 2013
Published: August 26, 2013

dx.doi.org/10.1021/ie3035306 | Ind. Eng. Chem. Res. 2013, 52, 13685−13698

development team, this approach tends to be preferentially used, as it is more flexible than the first class of approaches.

Finally, the third class of approaches is based on transforming the time-series of the original variables into other(s) sequence(s), presenting much less autocorrelation than the original one and, therefore, more easily monitored by the methodologies based on the iid assumption. A family of transforms with the ability to approximately diagonalize the autocorrelation function for a large class of signals is the wavelet transform family. The monitoring procedures thus obtained fall under the heading of multiscale statistical process control[22−24] and avoid extensive time-series modeling, leading to good results for a large class of process upsets and perturbation magnitudes, given their adaptive nature. However, their implementation requires a nontrivial programming task, a factor that could hinder their widespread use, unless deployed in proper software products, something that has not happened yet.

Among the three classes of approaches mentioned above, only the second and third show potential to be applied to the monitoring of large-scale dynamical processes. In this article, we focus on the second class of approaches, given the interest that it has been attracting and its familiarity to a large number of practitioners, an important point to guarantee the deployment and aimed practical impact of this research work.

When addressing the problem of monitoring large-scale processes with autocorrelation, the usual use of a preliminary stage of classical time-series modeling to generate monitoring residuals soon is discarded, as it very hard, if possible at all, to estimate a VARIMA (vector ARIMA) model even for low/moderate numbers of variables (such as 10). The solution usually proposed relies on employing latent-variable frameworks, which are well-known for their ability to handle a large number of correlated variables. These frameworks are then modified to incorporate the ability to also describe the autocorrelated structure of collected data. This is the basic reasoning underlying the dynamic principal component analysis (DPCA) methodology, proposed by Ku et al.[20] in 1995 in the scope of process monitoring, which has been employed as the standard technique for handling large-scale processes with autocorrelation (more information about DPCA will be provided in section 2).

However, even though DPCA constitutes an improved approach for conducting MSPC in large-scale dynamical systems when compared to the PCA-based methodology,[6,7,9] one can easily verify that the usual monitoring statistics, T^2 and Q (or SPE, squared prediction error), and the individual scores still present significant amounts of autocorrelation. Thus, such statistics still cannot be properly monitored with control charts based on iid assumptions, particularly when the amount of residual autocorrelation is relatively high.

The failure of generating decorrelated statistics in the original DPCA approach has a direct impact in the detection ability of the technique, as will be demonstrated in this article. This has led to the development of other approaches, more complex and involved, for handling the same monitoring problem. Examples of such methodologies include the use of state-space models[17,18,25] and the combination of state-space or multivariate time-series modeling frameworks with latent-variable models.[19,26,27] The relationship of these methodologies with the present work is addressed in the Discussion section.

In this article, we propose and analyze several approaches related to DPCA, some of them being new and having the ability to generate statistics with very low levels of autocorrelation. We compare their monitoring performances with a variety of current statistics, when applied to several case studies under different conditions. Among the approaches proposed, we found that the statistics derived from DPCA based on decorrelated residuals (DPCA-DR) consistently lead to superior detection performances and also have the advantages of being robust; easy to compute; structurally simpler; and more consistent, as the prediction task necessary to remove the dynamical effects from data is entirely accomplished within the same estimated DPCA model, instead of being carried out with resource to parallel modeling techniques, as in time-series or state-space models.

The rest of this article is organized as follows. In the next section, we present the core methodologies used in this study, with special focus on the latent-variable approaches employed (PCA, DPCA, and PLS). Then, we describe the current multivariate monitoring methods, as well as the prediction tools that they can integrate for properly describing the dynamic behavior of data. The core of this article lies in section 2, where we not only present the new methodologies, but do so from a global perspective that also incorporates the current methods as particular cases. In the following sections, we present and extensively discuss the results obtained. Finally, we discuss and summarize the main contributions presented in this article.

2. METHODS

In this section, we briefly present the latent-variable methodologies used in this work (mostly for the purpose of establishing the nomenclature), as well as the monitoring schemes that are tested, including the new methodologies proposed in this article.

2.1. Latent-Variable Methods. We begin by reviewing the latent-variable approaches employed in this work (namely, PCA, its extension DPCA, and PLS) and also address the methodology for estimating missing data in the context of PCA models, given its relationship to one of the steps of the proposed procedure for constructing decorrelated residuals.

2.1.1. Principal Component Analysis. Principal component analysis (PCA) is a multivariate data analysis technique focused on finding a low-dimensional subspace around which the majority of data variability is concentrated (called the PCA subspace). The new variables of such a low-dimensional subspace are linear combinations of the original variables and are called principal components (PCs). The PCs are uncorrelated quantities by design, with an implicit ordering, where the first few PCs concentrate the greatest portion of data variability. The remaining contributions of the subsequent components are gathered into a single residual matrix, **E**. Thus, PCA provides a decomposition of a data matrix, **X**, with n observations and m original variables, $\mathbf{X}_{n\times m}$, as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{T}_{n\times p}$ is the matrix of PCA scores (in which the ith column contains the scores for the ith PC), $\mathbf{P}_{m\times p}$ is the matrix with the PCA loadings (in which the ith column contains the variable loadings for the ith PC), and $\mathbf{E}_{n\times m}$ is the residual matrix. p represents the number of PCs retained (the data pseudorank), that is, the dimension of the PCA subspace.

2.1.2. Dynamic Principal Component Analysis. PCA describes only the correlations among variables, but does not incorporate any feature to address the correlations among the mode of observations, that is, the variable autocorrelation. Ku et al.[20] presented an approach for incorporating a linear time-series modeling framework into conventional PCA through a "time lag shift" methodology. It consists of adding several time-lagged replicates of the variables under analysis to the original set of variables and then applying PCA to the resulting extended matrix, say, $\tilde{\mathbf{X}}$, to also model the variables' dynamical structure, in addition to all of the static relationships. A possible notation for describing this computational scheme is given by

$$\tilde{\mathbf{X}} = \left[\overbrace{\mathbf{x}_1(0)\cdots\mathbf{x}_m(0)}^{\mathbf{x}(0)}\ \overbrace{\mathbf{x}_1(1)\cdots\mathbf{x}_m(1)}^{\mathbf{x}(1)}\cdots\overbrace{\mathbf{x}_1(l)\cdots\mathbf{x}_m(l)}^{\mathbf{x}(l)}\right] \tag{2}$$

where $\mathbf{x}_i(j)$ represents the ith variable (in column format) shifted j times into the past (i.e., with j lags); $\mathbf{x}(j)$ is the submatrix containing all the original variables shifted j times; and $\tilde{\mathbf{X}}$ is the resulting extended matrix (with l lags), which, in this case, has the form of a Hankel matrix.[28] However, in general, different lags can be used for different variables, in which case the structure of the extended matrix is no longer that of a Hankel matrix.

Therefore, in simple terms, DPCA is essentially the same method as the original PCA approach, except that the data matrix, $\tilde{\mathbf{X}}$, now includes additional time-shifted replicates of the original variables. Thus, after parameter estimation, it corresponds to an implicit vector autoregressive model[29] (VAR or VARX, if process inputs are also included; in fact, in more precise terms, the actual model structure corresponds to a latent-variable VAR or VARX). A key aspect when applying DPCA is the definition of its lag structure, for which some guidelines have already been proposed.[20,30]

After the construction of the extended set of process variables, $\tilde{\mathbf{X}}$, a PCA analysis can be carried out as usual, leading to the final DPCA model.

2.1.3. Partial Least Squares. Partial least squares (or projection onto latent structures, PLS) is a multivariate method that relates two data matrices, **X** and **Y**, through a linear latent-variable model.[31−34] More specifically, given an $(n \times m)$ data matrix of input variables, **X**, and an $(n \times r)$ data matrix with the corresponding output variables, **Y**, PLS successively finds those pairs of orthogonal linear combinations of the input and output variables with maximal covariance. If there is only one output variable, no linear combinations are considered in the **Y** block. In this way, PLS describes the part of the variability in **X** with predictive potential for explaining the variability of the outputs. The PLS model structure defines a common latent-variable space relating the two data blocks as follows[33,35,36]

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{3}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{4}$$

where $\mathbf{T}_{n\times p}$ are the **X** scores, defining the common latent-variable space relating $\mathbf{X}_{n\times m}$ and $\mathbf{Y}_{n\times r}$; $\mathbf{P}_{m\times p}$ and $\mathbf{Q}_{r\times p}$ are the **X**- and **Y**-loading matrices, respectively; $\mathbf{B}_{p\times p}$ is a regression coefficient matrix (for the inner latent-variable relationships); $\mathbf{E}_{n\times m}$ and $\mathbf{F}_{n\times r}$ are residual matrices; and p is the number of PLS components (i.e., latent-variables) considered.

As in the PCA case, a dynamic PLS model can be obtained by including past values of input and/or output variables in the **X** block.

2.1.4. Estimating Missing Data in PCA by Conditional Mean Replacement. Nelson et al. studied several approaches for estimating the PCA or PLS scores and responses when some observations are missing but a model is already available.[37] Among the approaches studied, conditional mean replacement was found to be, in general, the most favorable option, and it will be reviewed briefly here for the case of PCA.

Consider a data matrix **X** that can be decomposed by PCA as $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$, as in eq 3. When a new measurement vector is collected, say, $\mathbf{x} = [x_1\ x_2 \cdots x_m]^{\mathrm{T}}$, it might contain some missing measurements. For notational convenience and without loss of generality, we assume that such missing measurements are the first elements of this data vector, which then presents the following partitioned structure

$$\mathbf{x}^{\mathrm{T}} = [\mathbf{x}^{\#\mathrm{T}}\quad \mathbf{x}^{*\mathrm{T}}] \tag{5}$$

where $\mathbf{x}^{\#}$ denotes the missing measurements and $\mathbf{x}^{*}$ denotes the observed variables. Correspondingly, the loading matrix, **P**, can also be rearranged to conform to such a partition: $\mathbf{P}^{\mathrm{T}} = [\mathbf{P}^{\#\mathrm{T}}\ \mathbf{P}^{*\mathrm{T}}]$. The methodology for estimating the scores in this case is a particular case of the application of the expectation−maximization (EM) algorithm.[38] In general, this algorithm can be used to estimate a statistical model in the presence of missing data, $\mathbf{x}^{\#}$. In this approach, successive estimates of the missing data, $\hat{\mathbf{x}}^{\#}$, obtained with updated parameter estimates during the estimation stage are employed to refine the model parameters during the maximization stage. The successive refinements obtained for $\hat{\mathbf{x}}^{\#}$ correspond to the expected values of $\mathbf{x}^{\#}$, given the knowledge of the observed variables, $\mathbf{x}^{*}$, and the current estimates of the model parameters, $\boldsymbol{\theta}$:

$$\hat{\mathbf{x}}^{\#} = E(\mathbf{x}^{\#}|\mathbf{x}^{*}, \boldsymbol{\theta}) \tag{6}$$

In the present situation, it is assumed that a PCA model is already available (estimated from reference data), and therefore, only the expectation stage of the EM algorithm is necessary, to compute the expected values for missing measurements conditioned on the knowledge of $\mathbf{x}^{*}$ and the model parameters (PCA loadings and preprocessing parameters). We also consider the following notation for the spectral decomposition of the covariance matrix, $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}$, after introducing the above-mentioned missing data/observed variable partitioning

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^{\#}\boldsymbol{\Lambda}\mathbf{P}^{\#\mathrm{T}} & \mathbf{P}^{\#}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}} \\ \mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{\#\mathrm{T}} & \mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}} \end{bmatrix} \tag{7}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the PCA eigenvalues along the main diagonal and all other quantities have the same meanings as introduced before. Using this expression for $\boldsymbol{\Sigma}$, the conditional expectation of the missing measurements is simply given by

$$\hat{\mathbf{x}}^{\#} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}^{*} = \mathbf{P}^{\#}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}}(\mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}})^{-1}\mathbf{x}^{*} \tag{8}$$

The estimated missing measurements can then be used in the score calculation along with the observed data as if no measurements were missing, leading to the following expression for the first p scores (p is the pseudorank):

$$\begin{aligned} \hat{\mathbf{t}} &= \mathbf{P}_{1:p}{}^{\mathrm{T}}\begin{bmatrix} \hat{\mathbf{x}}^{\#} \\ \mathbf{x}^{*} \end{bmatrix} = \mathbf{P}_{1:p}{}^{\mathrm{T}}\begin{bmatrix} \mathbf{P}^{\#}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}}(\mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}})^{-1}\mathbf{x}^{*} \\ (\mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}})(\mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}})^{-1}\mathbf{x}^{*} \end{bmatrix} \\ &= \mathbf{P}_{1:p}{}^{\mathrm{T}}\begin{bmatrix} \mathbf{P}^{\#} \\ \mathbf{P}^{*} \end{bmatrix}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}}(\mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}})^{-1}\mathbf{x}^{*} \\ &= \mathbf{P}_{1:p}{}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}}(\mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}})^{-1}\mathbf{x}^{*} \\ &= [\mathbf{I}\quad \mathbf{0}]\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}}(\mathbf{P}^{*}\boldsymbol{\Lambda}\mathbf{P}^{*\mathrm{T}})^{-1}\mathbf{x}^{*} \end{aligned} \tag{9}$$

where $\mathbf{P}_{1:p}$ is composed of the first p columns of the loading matrix, $\mathbf{P}$; $\mathbf{I}$ is a $(p \times p)$ identity matrix, and $\mathbf{0}$ is a $[p \times (m - p)]$ matrix of zeros.

Even though we are not dealing with the missing data imputation problem in this article, we also apply this formalism to estimate the scores and current observations [i.e., $\mathbf{x}(0)$ in eq 2], given the knowledge of observations from the past and the reference DPCA model parameters. This will be instrumental for computing the decorrelated residuals and the new monitoring statistics with improved properties based on them.

On the other hand, by estimating the current observations conditioned on past observations and the estimated DPCA model, one is effectively employing the dynamical model structure for prediction purposes, through which one can effectively extract the dynamic trends of data variability and build more effective process monitoring schemes. Furthermore, this is achieved in a completely consistent and integrated way, without relying on a separate dynamical modeling task using additional time-series or state-space models.

2.2. Multivariate Statistical Process Control for Large-Scale Processes. In the previous section, we introduced the basic latent-variable models on which MSPC approaches for large-scale processes are based. The current methodologies for performing MSPC based on such procedures for static systems or when the variable autocorrelation is not very noticeable (PCA-MSPC and PLS-MSPC) are briefly described in Appendix A. In this section, we present monitoring procedures and corresponding statistics for handling situations in which the dynamic behavior of the process variables can no longer be overlooked. These statistics are introduced, for the first time, according to a general presentation framework that incorporates new monitoring approaches for large-scale systems with autocorrelation, together with the current ones. It is our belief that, using such a presentation scheme, the overall picture of the spectrum of approaches for dealing with this class of problems can be communicated more effectively. The monitoring statistics are based on a combination of approaches that are able to handle the high dimensionality and collinearity of process data (correlation) with others that take into account their dynamic features (autocorrelation). In particular, the approaches might encompass a subset of the following methodologies: PCA, DPCA, PLS, and ARMA (autoregressive moving average) models and missing data imputation methodologies.

Despite their variability, all of the methodologies mentioned in this section inherit some common features from the static approach already in use. In particular, they share the same latent-variable backbone for handling the correlated structure of data, leading to a decomposition of the overall data variability into two orthogonal parts, namely, the part captured and described by the model (followed by the Hotelling's T^2 of the scores) and the part not captured by it (residual variability, monitored with the SPE or Q statistic). As the several methodologies under consideration encompass multiple combinations of latent-variable models, lag-selection methods, prediction approaches, and monitoring subspaces, it becomes necessary to designate each monitoring statistics in an unambiguous, concise, and systematic way. To do so, a compound code was developed for defining which alternative for these different dimensions is actually being considered in a given monitoring statistic. More specifically, the dimensions specified are the latent-variable model employed (PCA, DPCA, DPLS) and, when required, the method used to select time-shifted variables (Table 1); the prediction method used to handle the autocorrelation structure of data; and the subspace the statistic will monitor (Table 2). The complete structure of this coding system is presented in the end of this section. We begin by introducing the several alternatives considered in each dimension.

Table 1. Definitions of the Codes Used to Identify the Lag-Selection Method Employed in DPCA

designation	lag-selection method
LS1	proposed by Ku et al.[20]
LS2	new[30]

Table 2. Codes Used to Identify the Part of Data Variability under Monitoring with a Given Statistic (S or R) and the Specific Type of Statistic Used (1−3)

latent-variable statistics (S)[a]		residual statistics (R)[b]	
code	statistic type	code	statistic type
S1	T^2 for observed scores	R1	Q (or SPE) obtained from reconstructed data using observed scores
S2	T^2 for observed and estimated scores	R2	T^2 for the residuals of reconstructed data using estimated scores
S3	T^2 for the difference between the observed and estimated scores (residual scores)		

[a] PCA subspace. [b] Original variables' subspace.

2.2.1. Code Dimension: Latent-Variable Model. The latent-variable model structures considered in this study are PCA, DPCA, and dynamic PLS (DPLS). For the DPCA approach, two different lag-selection (LS) methodologies were employed for defining its lag structure (i.e., the number of time-shifted replicates for each variable) (Table 1). The LS1 method was presented by Ku et al.[20] and is based on the number of linear relations required to properly describe the system. The LS2 method estimates the number of lags for each variable based on a succession of singular value decomposition problems and subsequent analyses of the results following a set of rules, and it is fully described elsewhere.[30]

2.2.2. Code Dimension: Prediction Method. In this article, we apply the term "observed scores" to the scores obtained directly from observed data (using a completely defined latent-variable model) through a direct projection operation and not estimated by some other mean, such as with resource to a time-series model (in the latent-variable space) or using a missing data imputation technique. If the scores are estimated by one of these approaches, they are called "estimated scores", and the estimation methodology is specified by this code dimension. The estimated scores $(\hat{\mathbf{t}})$ can be obtained either through the use of a time-series model (autoregressive) or by application of the conditional estimation method presented in section 2.1.4. For the cases in which an autoregressive (AR) model of order r is employed, the predicted score for the ith PC at the current time t, $\hat{t}_{i,t}$, is given by

$$\hat{t}_{i,t} = \nu_i + a_{i,1}t_{i,t-1} + a_{i,2}t_{i,t-2} + \cdots + a_{i,r}t_{i,t-r} \tag{10}$$

The appropriate AR model was fitted with resource to the Matlab algorithm ARfit developed by Schneider and Neumaier.[39] The AR model order was optimally selected

Table 3. Compound Coding Scheme Used to Designate All of the Monitoring Statistics Studied in This Work

class of latent-variable model	with/without time-shifted variables	lag-selection method	prediction model	designation of the statistic	equation	ref
PCA	without	–	–	PCA-0-S1	11	6
				PCA-0-R1	12	7
			time series	PCA-TS-S2	13[a]	new
				PCA-TS-S3	14[a]	–
				PCA-TS-R2	15[a]	new
	with	Ku et al.[20] (LS1)	–	DPCA-LS1-0-S1	11	20
				DPCA-LS1-0-R1	12	20
			time series	DPCA-LS1-TS-S2	13[a]	new
				DPCA-LS1-TS-S3	14[a]	–
				DPCA-LS1-TS-R2	15[a]	new
			missing data	DPCA-LS1-MD-S3	14[b]	new
				DPCA-LS1-MD-R2	15[b]	new
		proposed method (LS2)	–	DPCA-LS2-0-S1	11	–
				DPCA-LS2-0-R1	12	–
			time series	DPCA-LS2-TS-S2	13[a]	new
				DPCA-LS2-TS-S3	14[a]	–
				DPCA-LS2-TS-R2	15[a]	new
			missing data	DPCA-LS2-MD-S3	14[b]	new
				DPCA-LS2-MD-R2	15[b]	new
PLS	with	genetic algorithms (GA)[c]	–	DPLS-GA-S1	11	–
				DPLS-GA-R1x	12	–
				DPLS-GA-R1y	12	–

[a] Estimated scores ($\hat{\mathbf{t}}$) obtained through a time-series model, eq 10. [b] Estimated scores ($\hat{\mathbf{t}}$) obtained by conditional estimation, eq 9. [c] Time-sifted variables to be included in the *X*-block were determined by application of genetic algorithms on an extended data matrix.

using the Schwarz Bayesian criterion as the default method. If this was the prediction method used to estimate scores, the code "TS" (time-series) appears in the designation of the statistic. For situations in which conditional estimation was used, the respective code is "MD" (from its origin, missing data imputation theory). The corresponding statistics are also called DPCA-DR, as the residuals involved, $(\mathbf{t} - \hat{\mathbf{t}})$ and $(\mathbf{x} - \mathbf{P}\hat{\mathbf{t}})$, lead to statistics with very low levels of serial correlation (DR stands for decorrelated residuals). A code of "0" indicates that no predicted scores were used in the statistic.

2.2.3. Code Dimension: Monitored Subspace. Table 2 summarizes the part of the code used to refer to the different situations considered regarding the complementary partitioning of data variability. In this table, "reconstructed data" means reconstruction to the original variable domain of the variable values corresponding to the scores in the latent-variable space.

We now provide a more quantitative description of the statistics used in this study. The S1 and R1 statistics result from the direct application of Hotelling's T^2 and Q statistics formulas, respectively, in DPCA. They are defined by

$$\mathrm{S1} = \mathbf{t}^{\mathrm{T}}\mathbf{S}_{\mathbf{t}}^{-1}\mathbf{t} \tag{11}$$

$$\mathrm{R1} = (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \hat{\mathbf{x}}) \tag{12}$$

where, following the usual definition of these statistics, $\mathbf{t}_{p\times 1}$ is the vector of scores, $\mathbf{S}_{\mathbf{t}}$ is the sample covariance matrix of the scores, and $\hat{\mathbf{x}}$ is the reconstructed data using the observed scores. When the MSPC monitoring methodologies make use of a prediction framework, such as the one based on a time-series model (TS) or an estimation methodology for current data (MD), one can also obtain the one-step-ahead estimates for the current scores, $\hat{\mathbf{t}}_{p\times 1}$ (i.e., the values of the scores at the current time, using information from the past, until the last sampling time). Once available, these estimates can be treated in different ways. One approach is to incorporate these predicted scores and the observed ones in a Hotelling's T^2 statistic as follows (S2)

$$\mathrm{S2} = \begin{bmatrix} \mathbf{t} \\ \hat{\mathbf{t}} \end{bmatrix}^{\mathrm{T}} \mathbf{S}_{\mathbf{t},\hat{\mathbf{t}}}^{-1} \begin{bmatrix} \mathbf{t} \\ \hat{\mathbf{t}} \end{bmatrix} \tag{13}$$

where $\mathbf{S}_{\mathbf{t},\hat{\mathbf{t}}}$ is the sample covariance matrix of the combined scores. This augmented vector of observed ($\mathbf{t}$) and estimated ($\hat{\mathbf{t}}$) scores allows for the accommodation of some structured variability that the autoregressive models were not able to capture.

The R2 and S3 statistics are similar to S1 and R1 for the observed scores, but now involving only their one-step-ahead prediction counterparts ($\hat{\mathbf{t}}$), and are given by

$$\mathrm{S3} = (\mathbf{t} - \hat{\mathbf{t}})^{\mathrm{T}}\mathbf{S}_{\mathbf{t}-\hat{\mathbf{t}}}^{-1}(\mathbf{t} - \hat{\mathbf{t}}) \tag{14}$$

$$\mathrm{R2} = (\mathbf{x} - \mathbf{P}\hat{\mathbf{t}})^{\mathrm{T}}\mathbf{S}_{\mathbf{r}}^{-1}(\mathbf{x} - \mathbf{P}\hat{\mathbf{t}}) \tag{15}$$

where $\mathbf{S}_{\mathbf{t}-\hat{\mathbf{t}}}$ is the sample covariance matrix for the difference between the observed and estimated scores $(\mathbf{t} - \hat{\mathbf{t}})$ and $\mathbf{S}_{\mathbf{r}}$ is the sample covariance matrix of the residuals, in the original variable space, obtained from the reconstructed data using the estimated scores $(\hat{\mathbf{x}} = \mathbf{P}\hat{\mathbf{t}})$ and the actual measurements ($\mathbf{x}$).

Each one of these statistics, is completely specified by the compound code [latent-variable method]-[prediction method]-[type of statistic], where the field "latent-variable method" can contain PCA, DPCA-LS1, DPCA-LS2, or PLS (i.e., for the case of DPCA, one also specifies the lag-selection methodology used, LS1 or LS2); the "prediction method" can be TS, MD, or 0; and the "type of statistic", refers to the specific type of statistics used (see Table 2; in the case of PLS, the statistic R1 can be applied over the original domain of the *X* variables or *Y* variables, designated as R1x and R1y, respectively). For instance, the statistic with the designation DPCA-LS2-MD-S3 is computed according to the formula for S3 (it makes use of the scores estimated by missing data, MD) and is based on a DPCA model, where the number of lags was estimated by the LS2 method. Thus, the DPCA-DR statistics are all those sharing the code backbone DPCA-xx-MD-xx.

A summary of all of the studied statistics is presented in Table 3. In this table, the original reference for each statistic is also indicated, with new contributions from this work being labeled "New". For the statistics that are not entirely new because they are based on concepts already cited in the literature, the respective reference field was left blank.

3. RESULTS

In this section, we present the results of applying the monitoring statistics summarized in Table 3 to several simulated scenarios in which faults of different types and magnitudes are introduced, to accurately measure and compare their detection performance. The systems studied include the Wood and Berry column, the multivariate AR(1) process presented by Ku et al., [20] a large-scale process with 100 variables, and a continuous stirred-tank reactor (CSTR) system with a heating jacket under feedback control.

The use of simulated examples with different levels of complexity in this work is justified by the need to have, at this stage, precise measures for comparing all of the approaches considered, for which complete control over the exact moments of occurrence of the faults, their duration, location, and main characteristics (namely, magnitude) is required. Only under such conditions can the monitoring statistics be comparatively assessed in a rigorous way. After this stage, following the standard practice in this field and building on the results obtained, the most promising approaches are selected for application to real-world data sets and processes.

3.1. Scenario 1: Wood and Berry Column.

In this scenario, a model for the approximated dynamical behavior of a binary distillation column separating methanol from water is employed.[40] In this dynamic model, the methanol weight fractions in the distillate (x_D) and in the reboiler (x_B) (output variables) are expressed as functions of the steam flow rates of the reflux (F_R) and reboiler (F_S) (input variables). It consists of a linear dynamic model whose transfer function in the Laplace domain is given by

$$\begin{bmatrix} x_D(s) \\ x_B(s) \end{bmatrix} = \begin{bmatrix} \dfrac{12.8e^{-s}}{16.7s + 1} & \dfrac{-18.9e^{-3s}}{21s + 1} \\ \dfrac{6.6e^{-7s}}{10.9s + 1} & \dfrac{-19.4e^{-3s}}{14.4s + 1} \end{bmatrix} \begin{bmatrix} F_R(s) \\ F_S(s) \end{bmatrix} \tag{16}$$

During the simulations, F_R and F_S are assumed to be normally distributed with zero mean and unit variance. x_D and x_B are computed according to eq 16 with a random disturbance superimposed, given by the following transfer function related to the feed flow rate and feed composition[41]

$$\begin{bmatrix} x_{D,d}(s) \\ x_{B,d}(s) \end{bmatrix} = \begin{bmatrix} \dfrac{3.8e^{-8.1s}}{14.9s + 1} & \dfrac{0.22e^{-7.7s}}{21s + 1} \\ \dfrac{4.9e^{-3.4s}}{13.2s + 1} & \dfrac{0.14e^{-9.2s}}{12.1s + 1} \end{bmatrix} \begin{bmatrix} d_1(s) \\ d_2(s) \end{bmatrix} \tag{17}$$

where d_1 and d_2 follow a normal distribution with zero mean and with a variance set so that the added variability corresponds to approximately 10 dB, according to the following definition of signal-to-noise ratio (SNR)

$$\mathrm{SNR} = 10 \log_{10}\left[\frac{\mathrm{var}(x)}{\mathrm{var}(x_{\mathrm{noise}})}\right] \tag{18}$$

The observation measurement vector is defined as $\mathbf{x} = [x_D\ x_B\ F_R\ F_S]^T$. To construct the latent-variable models, 3000 observations were collected under normal operation conditions ($\mathbf{X}_{\mathrm{ref}}$). The data matrix $\mathbf{X}_{\mathrm{ref}}$ was then used to estimate the number of lags required for the construction of the DPCA model. From this analysis, the number of lags obtained through the use of the Ku et al.[20] approach (LS1) was two for all variables. On the other hand, the LS2 method led to a lag structure of $l = [2\ 2\ 9\ 4]$, that is, two additional time-shifted replicates for each x_D and x_B, nine for F_R, and four for F_S. It should be noted that the continuous-time model presented in eq 16, when sampled every minute, gives rise to the following discrete-time difference equations[42]

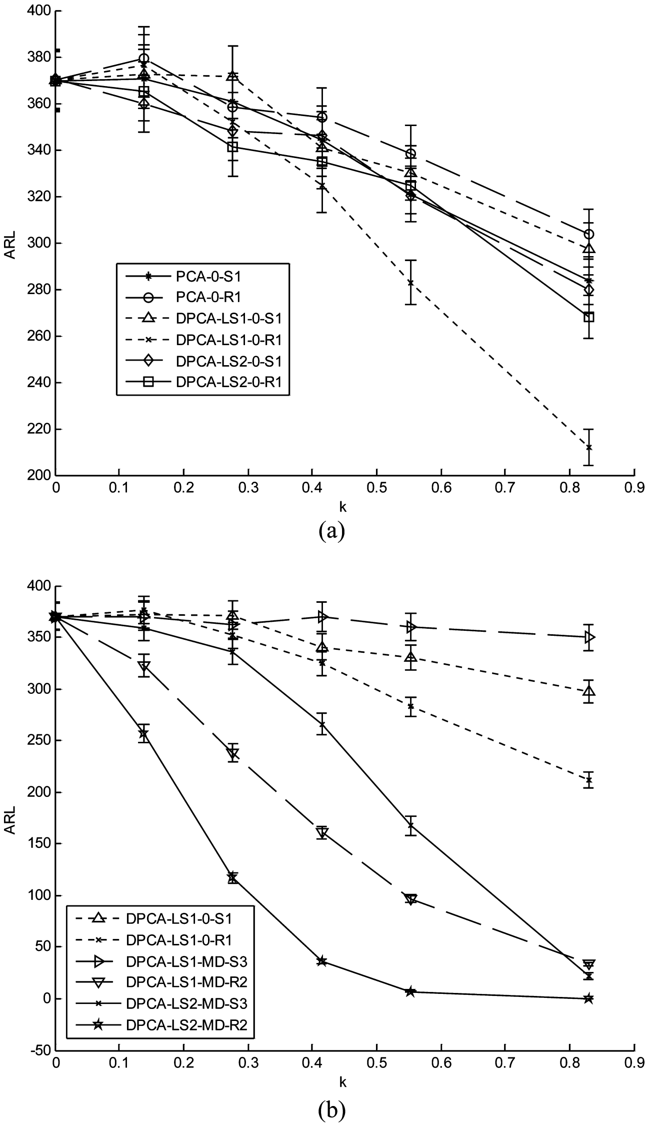
$$\begin{aligned} x_{D,i} &= 1.985x_{D,i-1} - 0.898x_{D,i-2} + 0.744F_{R,i-2} \\ &\quad - 0.709u_1F_{R,i-3} - 0.879F_{S,i-4} + 0.828F_{S,i-5} \\ x_{B,i} &= 184x_{B,i-1} - 0.851x_{B,i-2} + 0.579F_{R,i-8} - 0.54F_{R,i-9} \\ &\quad - 1.302F_{S,i-4} + 1.187F_{S,i-5} \end{aligned} \tag{19}$$

From these equations, one can verify that the actual lag structure is $l = [2\ 2\ 9\ 5]$. These numbers of lags, regarded as "true", are evidence of the superior lag-estimation ability of the LS2 method.

The system was then subjected to a set of step perturbations in the sensor measurements, and the corresponding average run length (ARL) was determined for each perturbation. In this process, the upper control limits (UCLs) for all statistics were previously adjusted, by trial and error, to enforce an equal in-control average run length (ARL_0) of 370 for each. This is a necessary procedure for enabling a sound comparison of all statistics, as it assures that they all present the same false alarm rates and, therefore, that the differences in fault detection performance arise only from their intrinsic characteristics and not from an arbitrary specification of the detection thresholds.

For each perturbation studied, 3000 data sets were generated, leading to the computation of 3000 run lengths, from which the ARLs were computed for all statistics. The ARL values (along with their associated 95% confidence intervals, obtained by bootstrapping) for a step perturbation in the mean of the first sensor, with magnitude k times the variable's standard deviation, are presented in Figure 1, where it can be seen that, in this case, there is no significant difference between the static and the dynamic versions of PCA (Figure 1a), even with the use of the LS2 method to estimate the number of lags. In fact, DPCA-LS1-0-R1 (which uses the Ku et al.[20] approach) gives better results than DPCA-LS2-0-R1. However, even with DPCA, the resulting statistics still present some autocorrelation (see Figure 2), and therefore, in Figure 1b, we present the results obtained with methodologies that incorporate an implicit prediction methodology, namely, through the conditional estimation approach (MD). For comparison purposes, we also present the results for the best monitoring statistic found in Figure 1a. The results obtained clearly show that the application of such an approach not only reduces the statistics autocorrelation (Figure 3), but also improves the control chart performance (Figure 1b).

In Figure 4, a performance comparison index (N) based on the computation of the area under the ARL curves (such as



(a)

(b)

Figure 1. ARLs for the tested methodologies used to monitor the Wood and Berry distillation column subject to deviations in the first sensor measurement. (a) No predictive methodologies employed, (b) approaches based on MD employed.

those shown in Figure 1) is presented. This index is normalized so that it falls in the range [0, 1], where 1 represents the best performance (smallest area under the ARL curve). In this analysis, the missing-data-based statistics, especially DPCA-LS2-MD-R2, present the best performance and lead to the weakest final autocorrelation.

3.2. Scenario 2: Multivariate AR(1) Process.

The following multivariate AR(1) process was presented by Ku et al.[20] to demonstrate the application of DPCA to multivariate statistical process control

$$\mathbf{z}(k) = \begin{bmatrix} 0.118 & -0.191 \\ 0.847 & 0.264 \end{bmatrix}\mathbf{z}(k-1) + \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}\mathbf{u}(k-1)$$
$$\mathbf{y}(k) = \mathbf{z}(k) + \mathbf{v}(k) \tag{20}$$

where **u** is the correlated input

$$\mathbf{u}(k) = \begin{bmatrix} 0.811 & -0.226 \\ 0.477 & 0.415 \end{bmatrix}\mathbf{u}(k-1) + \begin{bmatrix} 0.193 & 0.689 \\ -0.320 & -0.749 \end{bmatrix}\mathbf{w}(k-1) \tag{21}$$

The input **w** is a random noise sequence with zero mean and variance 1. The output **y** is equal to **z** plus another random noise component, **v**(k), with zero mean and variance 0.1. The observation measurement vector is defined as $\mathbf{x} = [\mathbf{y}^{\mathrm{T}}\ \mathbf{u}^{\mathrm{T}}]^{\mathrm{T}}$.

As in the previous examples, the reference data were composed of 3000 observations and used to estimate the reference models and the number of lags according to the two considered methodologies. In this case, both lag-selection methods gave an estimation of one lag for all variables, and therefore, there are no differences between statistics that differ in terms of only LS1 and LS2.

To assess the performance of the statistics, the system was subjected to changes in the mean of **w**, from which 3000 observations were collected for different magnitudes of the change. Each of these changes was repeated 3000 times. The corresponding ARL values for the more relevant statistics and their 95% confidence limits are presented in Figure 5.

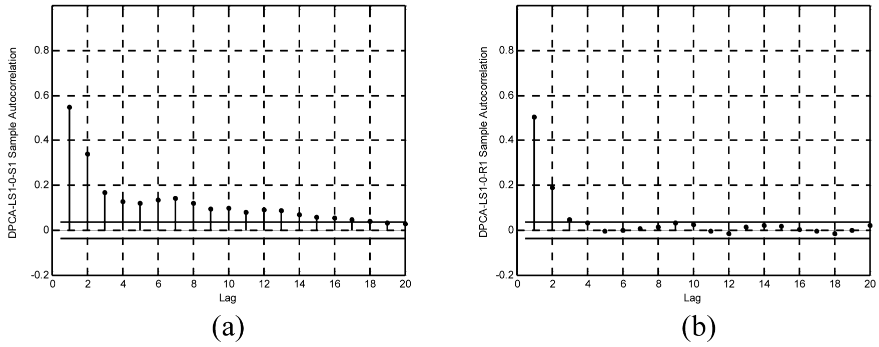
From Figure 5, it is noticeable that the proposed missing-data-based statistics present a superior performance, especially, DPCA-LS2-MD-R2, which also presents no residual autocorrelation, in contrast to what happens with the other monitoring approaches.

3.3. Scenario 3: Large-Scale Process.

To assess the ability of the proposed methodology to handle a large number of variables, a megavariate process with the following latent-variable model structure was simulated

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{22}$$

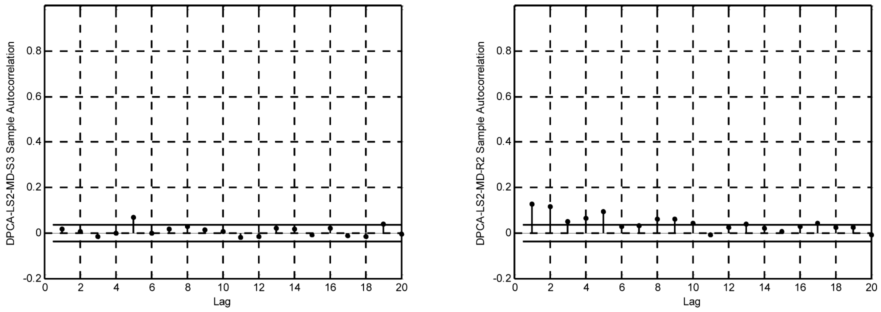
where **X** is an ($n \times m$) matrix of measured variables, **T** is an ($n \times p$) matrix of latent variables, **P** is an ($m \times p$) matrix of



(a)

(b)

Figure 2. Sample autocorrelation functions for the DPCA-LS1-0-S1 and DPCA-LS1-0-R1 statistics.



(a) (b)

Figure 3. Sample autocorrelation functions for the DPCA-LS2-MD-S3 and DPCA-LS2-MD-R2 statistics.

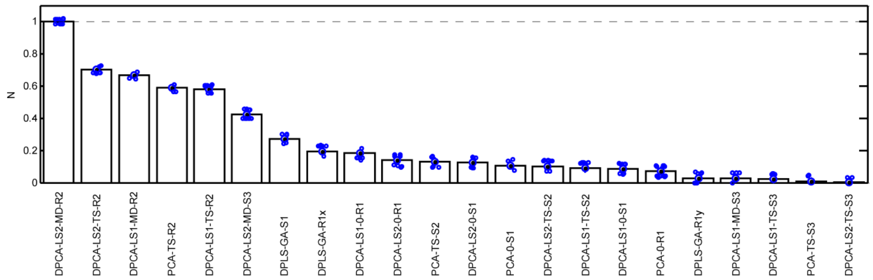


Figure 4. Comparison of the performance of different methods in the wood and berry column: Performance index (N) obtained when the system was subjected to deviations in the first sensor measurement. The heights of the bars correspond to the associated mean values.

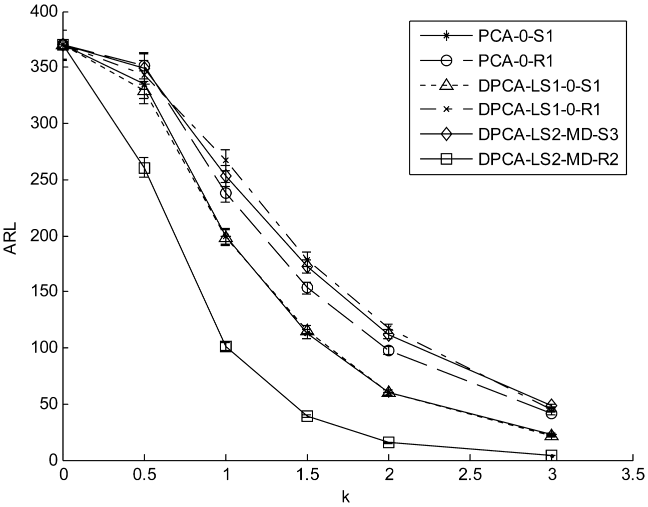


Figure 5. ARLs for the tested methodologies used to monitor the multivariate AR(1) process subject to step perturbations in the mean of **w**, with a magnitude of k times the standard deviation of **w**.

orthogonal loadings, and **E** is an $(n \times m)$ matrix of errors. In this study, the number of measured variables (m) was set to 100, with 5 latent variables (p) following independent AR(1) processes with autoregression coefficients of 0.90. The **P** matrix was randomly generated but forced to have orthogonal columns.

As a result of the application of the LS1 lag-selection method, zero lags were attributed to all variables, which is clearly an underestimation of the system's dynamics. Furthermore, this situation corresponds to the standard PCA procedure, and therefore, in this case, it was decided to use one lag for all variables to have a more diversified set of monitoring methodologies to compare. In contrast, when the LS2 lag-selection method was applied, it selected zero lags for 30 variables, one lag for 32 variables, and two lags for the remaining 38 variables. After estimating the reference models from data obtained under normal operation conditions (NOCs), the control limits for the monitoring statistics were adjusted to the same in-control ARL of 370. The system was then subjected to step deviations of different magnitudes in one of the variables of the **X** matrix, emulating, for instance, a sensor failure. The corresponding ARL for each perturbation was computed based on 3000 replications (leading to 3000 run lengths), which were represented along with the associated 95% confidence limits obtained by bootstrapping. Figure 6 illustrates the results obtained for a selected set of the most relevant monitoring statistics in this case study.

From Figure 6, it is clear that, in general, the monitoring methodologies are capable of detecting the simulated faults through their residual statistics (i.e., PCA-0-R1, DPCA-LS1-0-R1, and DPCA-LS2-MD-R2). However, under normal operation conditions, some of these procedures present a considerable autocorrelation, as can be seen in Figures 7–9. The effects of autocorrelation are more evident in the monitoring statistics for the scores subspace and to a lesser extent in the DPCA-LS1-0-R1 statistic applied to the residuals subspace, which implies that its underlying monitoring latent-variable models do not fully describe the system structure. It is also noticeable that the residual monitoring statistic of the static PCA model (Figure 7) has no autocorrelation, which explains its good detection capabilities in this case. Yet, the PCA monitoring statistic for the scores subspace is highly autocorrelated. On the other hand, when the decorrelated residuals approach is considered (DPCA-LS2-MD-S3 and DPCA-LS2-MD-R3), the monitoring statistics' autocorrelation remains low at all times (Figure 9) without compromising the detection ability.

From the performance comparison index (N) presented in Figure 10, it is also clear that the best monitoring statistics for

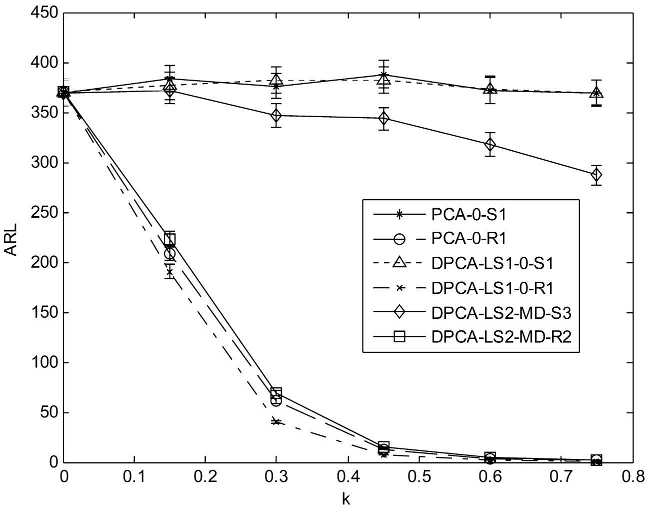


Figure 6. ARLs for the tested methodologies used to monitor the large-scale process subject to a step perturbation in one of the *X* variables, with a magnitude of *k* times the standard deviation of the corresponding variable.

this case study use either a time-series or a decorrelated-residuals approach to estimate the scores. However, we consider the latter to be a preferable approach, because it does not require the use of a different model structure other than DPCA, namely, a different time-series model for each score. Finally, even though DPCA-LS2-MD-R3 is, in this particular case study, the third best monitoring statistic, its performance is very close to that of the best statistic, as can be seen in Figure 6, and it has the advantage of using a more consistent procedure with low autocorrelation for both monitoring statistics (see Figure 9), which provides more robustness and simplicity in practice.

3.4. Scenario 4: Continuous Stirred-Tank Reactor. To assess the statistics performance in a more realistic system, a dynamical model of a continuous stirred-tank reactor (CSTR) with a heating jacket and under feedback control was employed. In this system, an endothermic reaction of the type A → B takes place in the CSTR with free discharge flow. This system is under a proportional–integral (PI) control system to maintain the temperature and fluid level close to their set points. The inputs of the system are the concentration and temperature of the feed stream and the inlet temperature of the heating fluid. The system outputs are the CSTR level, concentrations, and temperature and the heating-fluid outlet temperature. These three input and four output variables were considered to form the set of measured variables.

The method's parameters were determined from a reference data set composed of 3000 observations, leading to two lags for all of the mentioned variables trough the LS1 method. On the other hand, the LS2 method estimated only three lags for the reactor temperature and four lags for the heating-fluid outlet temperature.

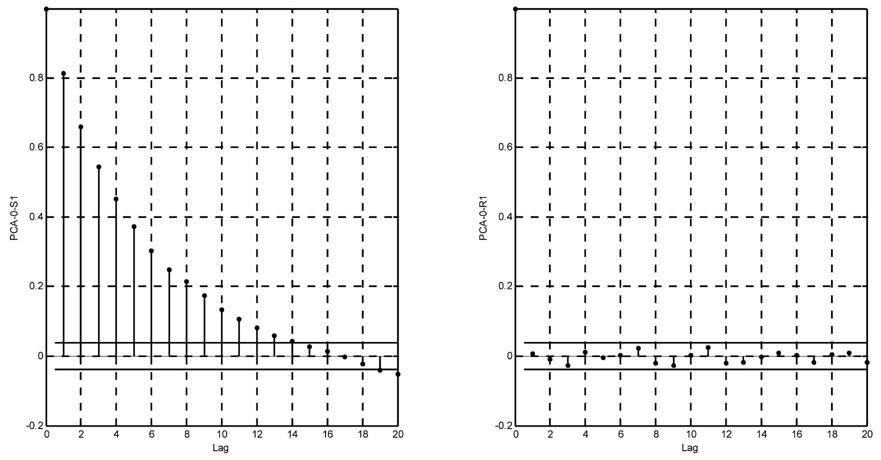
The system was then subjected to perturbations in the discharge coefficient and heat-transfer coefficient, which inherently change the systems dynamics and, therefore, affect several variables simultaneously. Each perturbation was repeated 3000 times, and in each run, 3000 observations were recorded.

In this case study, most of the statistics presented similar performances, as can be seen in Figures 11 and 12, where the results for the performance comparison index (*N*) are presented. Nevertheless, the inclusion of MD techniques eliminated the autocorrelation when applied with LS1, which was still outperformed by DPCA-LS2-MD-R2.

4. DISCUSSION

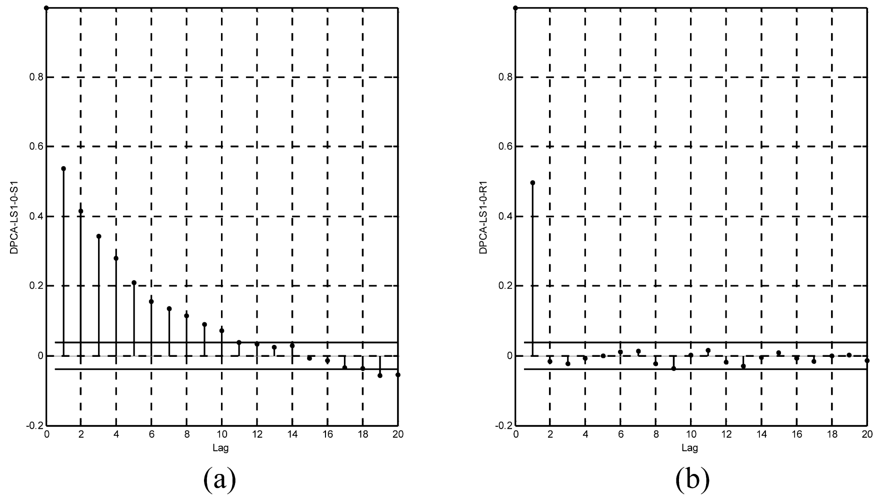
Analyzing the results presented in the previous section, one can verify that, consistently, the DPCA-LS2-MD-R2 statistic tended to present superior performances compared the remaining methodologies (Figure 13).

The DPCA-LS2-MD-R2 statistic was one of the new statistics introduced in this article and is representative of a new class of statistics that make use of data imputation methods to predict the current scores of a DPCA model, called DPCA-DR. Our results indicate that this class of statistics provides a competitive alternative to the current MSPC methodologies, as they usually present better detection performances and lower autocorrelation levels. However, they also depend on a suitable method for estimating the number of lags for the DPCA model. This issue was also addressed in this article, where two methods were employed to estimate the number of lags (one of them also new), and it was found that the LS2 method was clearly superior to LS1.



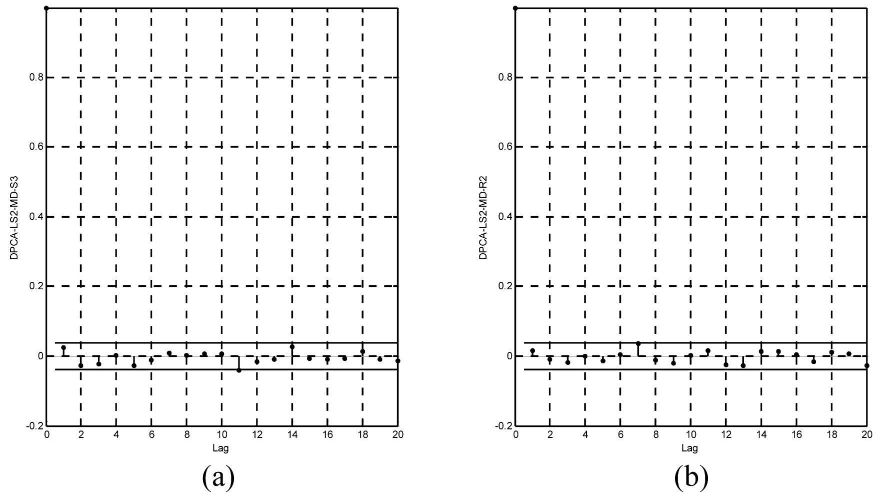
(a) (b)

Figure 7. Sample autocorrelation functions for the (a) PCA-0-S1 and (b) PCA-0-R1 statistics.



(a) (b)

Figure 8. Sample autocorrelation functions for the (a) DPCA-LS1-0-S1 and (b) DPCA-LS1-0-R1 statistics.



(a) (b)

Figure 9. Sample autocorrelation functions for the (a) DPCA-LS2-MD-S3 and (b) DPCA-LS2-MD-R2 statistics.

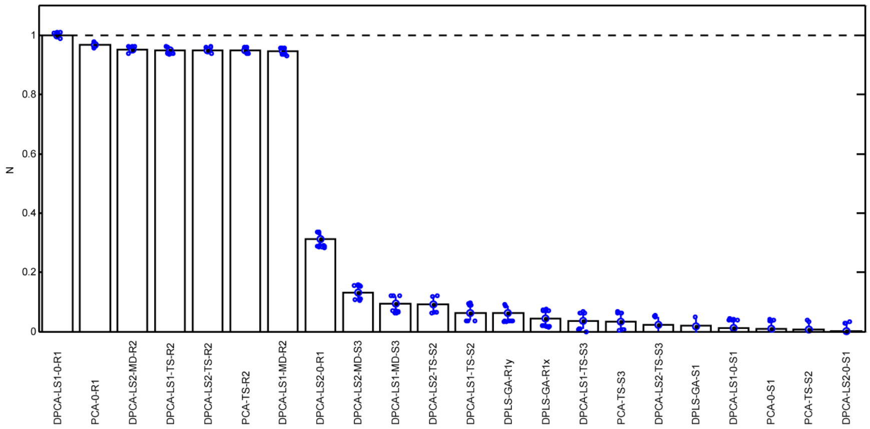


Figure 10. Comparison of the performance of different methods in the large-scale system: Performance index (N) obtained when the system was subjected to a step perturbations in one of the X variables. The heights of the bars correspond to the associated mean values.

The DPCA-DR approach implicitly conducts a multivariate time-series modeling and prediction to compute the residuals in the score and variables spaces but in a fully integrated way in a conventional DPCA framework, avoiding the usual two-stage modeling/projection sequence. Until now, applications of the DPCA model have not considered the computation of serial decorrelated residuals,[19,20,25,43] and we believed that this was limiting the performance and, therefore, the application of this approach in practice. For instance, in the work of Russel et al.,[20] DPCA was implemented as suggested by Ku et al.,[20] and therefore, no decorrelating residuals were computed.[25] This justifies the similar detection performances obtained with the

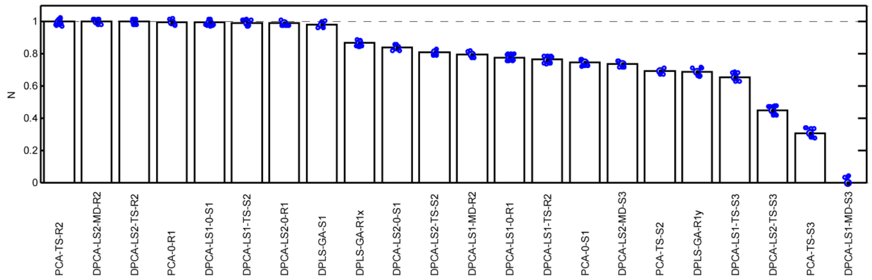


Figure 11. Comparison of the performance of the different methods in the CSTR system: Performance index (N) obtained when the system was subjected to changes in the discharge coefficient. The heights of the bars correspond to the associated mean values.

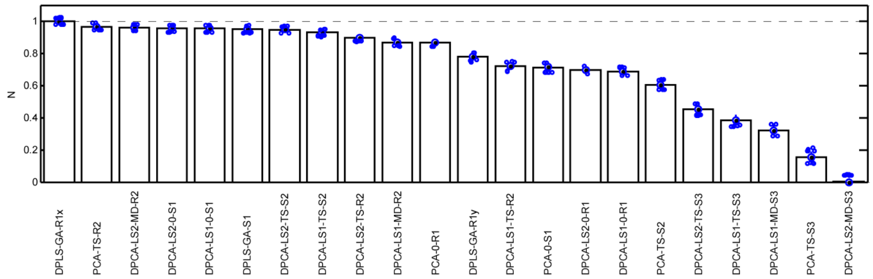


Figure 12. Comparison of the performance of the different methods in the CSTR system: Performance index (N) obtained when the system was subjected to changes in the heat-transfer coefficient. The heights of the bars correspond to the associated mean values.

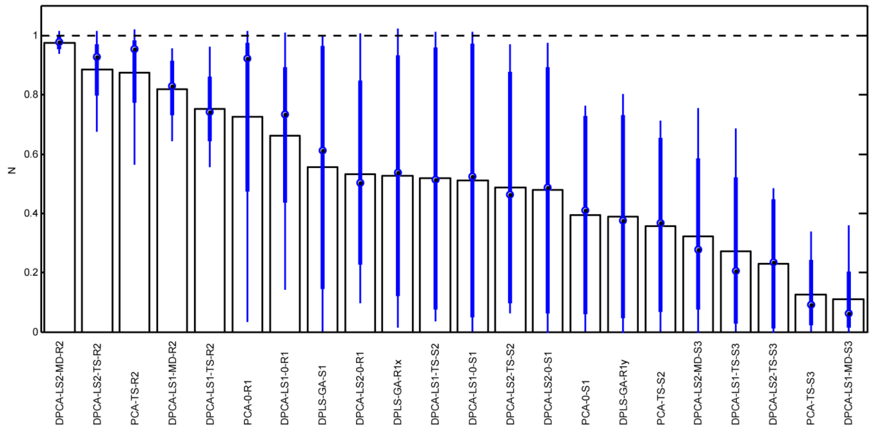


Figure 13. Comparison of the performance of the different methods in all systems considered in this study: Performance index (N) obtained in all simulations performed. The heights of the bars correspond to the associated mean values.

DPCA and PCA statistics in this work, as well as the inferior sensitivity of DPCA when compared with canonical variate analysis (CVA), which is also a linear modeling approach for estimating a state-space model. With the results proposed in our work, the performance of DPCA is significantly improved. Ongoing research indicates, for instance, that the proposed statistic for monitoring the variability around the principal-component subspace (usually the most sensitive statistic) is substantially more sensitive that the corresponding CVA Q statistics when applied to the same case study.[30] Therefore, when compared to linear state-space methodologies, an approach based on a DPCA framework not only has the advantage of being more familiar to practitioners (an important driver for its application in practice), but also presents the important features of being more robust (for instance, one of the CVA statistics proposed by Russel et al.[25] is very sensitive because of the inversion of a matrix with small values), simpler (a CVA model is more complex to identify and implement,[44] and the method requires three monitoring statistics;[25] it might also require just one statistic to evaluate if the process variable remains in the same region of the state-space, but then its application is restricted to the case where the states follow a multivariate Gaussian distribution[17,18]), and finally presents very competitive monitoring performances.

On the other hand, state-space and time-series approaches are focused on predicting the future system responses and might not properly describe the normal behavior of process inputs,[45] which gives rise to problems in process monitoring applications. To circumvent this problem, Treasure et al. proposed a combination of state-space models, identified with a subspace system identification methodology,[46] and PCA to incorporate both the description of the variability arising from the process dynamical behavior and that for the process inputs. Other combined approaches can also be found elsewhere.[26,27] However, compared to the approach proposed in the present work, two linear modeling frameworks are being applied in these solutions, instead of one, indicating higher complexity and implementation burden, even though they are still combining just two linear approaches. On the other hand, dynamic PLS[21,47,48] can also be used for estimating the underlying state-space model, with both state-space compression and partial consideration of the variations of the input variables in the model, all within a single linear modeling framework. For this reason, we used this methodology as a benchmark method in this work, instead of CVA and other subspace system identification methodologies [e.g., N4SID (numerical algorithm for subspace state space system identification)] and combined methods.

A possible limitation of the DPCA-DR approach is that it is based on the conditional mean replacement solution to the missing data imputation problem, which faces problems when the matrix $\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*$ becomes ill-conditioned. In such case, traditional regression solutions such as ridge regression, principal component regression, and PLS can be used. Alternatively, one can also use the projection onto the model plane approach[37] or trimmed scores regression (TSR)[49] to estimate the scores corresponding to missing observations.

5. CONCLUSIONS

In this work, we have addressed the problem of monitoring large processes with autocorrelated or dynamical data. Twenty-two monitoring statistics were presented and studied in a systematic way, including 10 statistics introduced for the first time. They encompass a variety of methods, including PCA, DPCA, PLS, time-series, and conditional estimation frameworks, and we found that, in general, those derived from the class of DPCA-DR statistics tend to show better detection performances. Such statistics require a proper method to estimate the number of lags necessary to construct the DPCA model, for which two alternatives were considered. From the consistency of the results obtained, we believe that the new proposed statistics are indeed valid alternatives to the current approaches strictly based on PCA and DPCA.

Future work will address the application of the most promising approaches identified in this work to real-world industrial data, now that these approaches have demonstrated their potential under challenging simulation scenarios.

■ APPENDIX A. CURRENT MULTIVARIATE STATISTICAL PROCESS CONTROL METHODS FOR LARGE-SCALE PROCESSES

Current approaches for implementing large-scale process monitoring are essentially based on latent-variable frameworks, given their intrinsic ability to deal with a high number of correlated variables. In fact, the classical full-rank MSPC approach, based on the Hotelling's T^2 statistic, is only rarely applied to processes with more than a dozen of variables, leaving outside its scope the vast majority of current multivariate process applications.

A common procedure for handling highly collinear process data in MSPC consists of using the significant PCA scores as the new variables for monitoring. As these variables are few in number and uncorrelated, the Hotelling's T^2 procedure can be applied without any limitation to the first p PCs.[6,7] In this case, the following monitoring statistic is applied, after all variables have been centered to zero mean, and could also be properly scaled

$$T_{\mathrm{PCA}}^{\ 2} = \sum_{i=1}^{p} \frac{t_i^{\ 2}}{\lambda_i} = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{\Lambda}_p^{\ -1}\mathbf{P}^{\mathrm{T}}\mathbf{x} \tag{23}$$

where $\mathbf{\Lambda}_p$ is a diagonal matrix with the first p eigenvalues in the main diagonal (the eigenvalue associated with a given PC also provides the value for its variance in the data set used to estimate the model). If the process under normal operation conditions (NOCs) follows a multivariate normal distribution, then the upper control limit (UCL) for $T_{\mathrm{PCA}}^{\ 2}$ is given by[11,50]

$$\mathrm{UCL} = \frac{p(n-1)(n+1)}{n^2 - np} F_{\alpha,p,n-p} \tag{24}$$

where $F_{\alpha,p,n-p}$ is the upper αth percentile of the F distribution, with p and $n-p$ degrees of freedom.

However, process monitoring strictly on the PCA subspace using $T_{\mathrm{PCA}}^{\ 2}$ based on the first p PCs is not sufficient, as it lacks a very important piece of information arising from the variability around the PCA subspace. This variability is captured by a monitoring statistic based on the squared prediction error (SPE) of the observations residuals, $\mathbf{e}_{m\times 1}$.[7] This statistic is also known as the Q statistic

$$Q = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \hat{\mathbf{x}}) = \mathbf{x}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}\mathbf{P}^{\mathrm{T}})\mathbf{x} \tag{25}$$

where $\hat{\mathbf{x}}$ stands for the projection of $\mathbf{x}$ onto the PCA subspace (i.e., the reconstruction in the original variable space) of the score "observed" in the latent-variable subspace. The process is considered to be under statistical control, if this statistic is below the upper control limit[7,51]

$$\mathrm{UCL} = \theta_1\left(\frac{c_\alpha\sqrt{2\theta_2 h_0^{\ 2}}}{\theta_1} + 1 + \frac{\theta_2 h_0(h_0 - 1)}{\theta_1^{\ 2}}\right)^{1/h_0} \tag{26}$$

where

$$\theta_i = \sum_{j=p+1}^{m} \lambda_j^i, \qquad i = 1, 2, 3 \tag{27}$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^{\ 2}} \tag{28}$$

where p is the number of retained principal components and c_α is the standard normal variable, corresponding to the upper $(1 - \alpha)$th percentile.

A similar MSPC procedure can also be derived for PLS, using the two orthogonal components of variability, namely, the variability in the PLS predictive subspace (PLS $\mathbf{X}$ scores) and the residuals around such subspace, or those for the response variable(s).[8,9,11] In this case, the Hotelling's T^2 statistic for a new score vector, $\mathbf{t}_{p\times 1}$, is given by[52]

$$T_{\mathrm{PLS}}^{\ 2} = \mathbf{t}^{\mathrm{T}}\mathbf{S}_{\mathbf{t}}^{\ -1}\mathbf{t} \tag{29}$$

where $\mathbf{S}_{\mathbf{t}}$ represents the estimated covariance matrix of the PLS $\mathbf{X}$ scores. The upper control limit (UCL) for $T_{\mathrm{PLS}}^{\ 2}$ is given by

$$\mathrm{UCL} = \frac{p(n-1)(n+1)}{n^2 - np} F_{\alpha,p,n-p} \tag{30}$$

where p is the number of latent variables retained in the PLS model and $F_{\alpha,p,n-p}$ is the upper αth percentile of the F distribution with p and $n - p$ degrees of freedom. As to the SPE (or Q statistic), it can be either computed for the $\mathbf{X}$ or the $\mathbf{Y}$ block, according to the following expressions:[9,10]

$$Q_{\mathbf{X}} = (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \hat{\mathbf{x}}) \tag{31}$$

$$Q_{\mathbf{Y}} = (\mathbf{y} - \hat{\mathbf{y}})^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{y}}) \tag{32}$$

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are the reconstructed variables in the original **X** and **Y** domains, respectively, using the PLS model defined in eqs 3 and 4.

The control limits for the Q_X and Q_Y statistics, can be determined assuming an approximation by a $g\chi^2$ distribution: UCL = $g\chi_{\alpha,h}{}^2$.[52,53] This equation is also well approximated by:[52]

$$\mathrm{UCL} = gh\left[1 - \frac{2}{9h} + c_\alpha\left(\frac{2}{9h}\right)^{1/2}\right]^3 \tag{33}$$

where g is a weighting factor and h the effective number of degrees of freedom for the χ^2 distribution, which can easily be obtained by matching the moments of these distribution with those from the empirical distribution, leading to $g = v/(2m)$ and $h = 2m^{2/}v$, where v is the variance and m is the mean of the SPE values (Q_X and Q_Y); c_α is the standard normal variable, corresponding to the upper $(1 - \alpha)$th percentile.

AUTHOR INFORMATION

Corresponding Author

*E-mail: marco@eq.uc.pt. Tel.: +351 239 798 700. Fax: +351 239 798 703.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

T.J.R. acknowledges the Portuguese Foundation for Science and Technology for his Ph.D. grant (Grant SFRH/BD/65794/2009). M.S.R. also acknowledges financial support through project PTDC/EQU-ESI/108374/2008 cofinanced by the Portuguese FCT and European Union's FEDER through "Eixo I do Programa Operacional Factores de Competitividade (POFC)" of QREN (FCOMP-01-0124-FEDER-010397).

REFERENCES

(1) Shewhart, W. A. *Economic Control of Quality of Manufactured Product*; 50th Anniversary Commemorative Reissue; ASQC Quality Press: Milwaukee, WI, 1980.

(2) Roberts, S. W. Control Charts Tests Based on Geometric Moving Averages. *Technometrics* **1959**, *1* (3), 239−250.

(3) Page, E. S. Continuous Inspection Schemes. *Biometrics* **1954**, *41* (1−2), 100−115.

(4) Hotelling, H. Multivariate quality control, illustrated by the air testing of sample bombsights. In *Selected Techniques of Statistical Analysis*; Eisenhart, C., Hastay, M. W., Wallis, W. A., Eds.; McGraw-Hill: New York, 1947.

(5) Lowry, C. A.; Woodall, W. H.; Champ, C. W.; Rigdon, S. E. A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics* **1992**, *34* (1), 46−53.

(6) Jackson, J. E. Quality Control Methods for Several Related Variables. *Technometrics* **1959**, *1* (4), 359−377.

(7) Jackson, J. E.; Mudholkar, G. S. Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics* **1979**, *21* (3), 341−349.

(8) Kourti, T.; MacGregor, J. F. Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods. *Chemom. Intell. Lab. Syst.* **1995**, *28*, 3−21.

(9) Kresta, J. V.; MacGregor, J. F.; Marlin, T. E. Multivariate Statistical Monitoring of Process Operating Performance. *Can. J. Chem. Eng.* **1991**, *69*, 35−47.

(10) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE J.* **1994**, *40*, 5.

(11) MacGregor, J. F.; Kourti, T. Statistical Process Control of Multivariate Processes. *Control Eng. Pract.* **1995**, *3* (3), 403−414.

(12) Vasilopoulos, A. V.; Stamboulis, A. P. Modification of Control Chart Limits in the Presence of Data Correlation. *J. Qual. Technol.* **1978**, *10* (1), 20−30.

(13) Lu, C.-W.; Reynolds, M. R., Jr. Cusum Charts For Monitoring An Autocorrelated Process. *J. Qual. Technol.* **2001**, *33* (3), 316−334.

(14) Vermaat, M. B.; Does, R. J. M. M.; Bisgaard, S. EWMA Control Chart Limits for First- and Second-Order Autoregressive Processes. *Qual. Reliab. Eng. Int.* **2008**, *24*, 573−584.

(15) Harris, T. J.; Ross, W. H. Statistical Process Control Procedures for Correlated Observations. *Can. J. Chem. Eng.* **1991**, *69*, 48−57.

(16) Montgomery, D. C.; Mastrangelo, C. M. Some Statistical Process Control Methods for Autocorrelated Data. *J. Qual. Technol.* **1991**, *23* (3), 179−193.

(17) Negiz, A.; Çinar, A. Statistical Monitoring of Multivariable Dynamic Processes with State-Space Models. *AIChE J.* **1997**, *43* (8), 2002−2020.

(18) Simoglou, A.; Martin, E. B.; Morris, A. J. Dynamic Multivariate Statistical Process Control Using Partial Least Squares and Canonical Variate Analysis. *Comput. Chem. Eng.* **1999**, *23* (Suppl.), S277−S280.

(19) Treasure, R. J.; Kruger, U.; Cooper, J. E. Dynamic Multivariate Statistical Process Control Using Subspace Identification. *J. Process Control* **2004**, *14*, 279−292.

(20) Ku, W.; Storer, R. H.; Georgakis, C. Disturbance Detection and Isolation by Dynamic Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 179−196.

(21) Komulainen, T.; Sourander, M.; Jämsä-Jounela, S.-L. An Online Application of Dynamic PLS to a Dearomatization Process. *Comput. Chem. Eng.* **2004**, *28*, 2611−2619.

(22) Bakshi, B. R. Multiscale PCA with Application to Multivariate Statistical Process Control. *AIChE J.* **1998**, *44* (7), 1596−1610.

(23) Reis, M. S.; Bakshi, B. R.; Saraiva, P. M. Multiscale Statistical Process Control Using Wavelet Packets. *AIChE J.* **2008**, *54* (9), 2366−2378.

(24) Reis, M. S.; Saraiva, P. M. Multiscale Statistical Process Control with Multiresolution Data. *AIChE J.* **2006**, *52* (6), 2107−2119.

(25) Russel, E. L.; Chiang, L. H.; Braatz, R. D. Fault Detection in Industrial Processes Using Canonical Variate Analysis and Dynamic Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 81−93.

(26) Kaspar, M. H.; Ray, W. H. Dynamic PLS Modelling for Process Control. *Chem. Eng. Sci.* **1993**, *48* (20), 3447−3461.

(27) Lakshminarayanan, S.; Shah, S. L.; Nandakumar, K. Modeling and Control of Multivariable Processes: Dynamic PLS Approach. *AIChE J.* **1997**, *43* (9), 2307−2322.

(28) Ljung, L. *System Identification: Theory for the User*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 1999.

(29) Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C. *Time Series Analysis: Forecasting and Control*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, 1994.

(30) Rato, T. J.; Reis, M. S. Defining the Structure of DPCA Models and Its Impact on Process Monitoring and Prediction Activities. *Chemom. Intell. Lab. Syst.* **2013**, *125* (15), 74−86.

(31) Helland, I. S. On the Structure of Partial Least Squares Regression. *Commun. Stat.-Simul.* **1988**, *17* (2), 581.

(32) Höskuldsson, A. PLS Regression Methods. *J. Chemom.* **1988**, *2*, 211−228.

(33) Martens, H.; Naes, T. *Multivariate Calibration*. Wiley: Chichester, U.K., 1989.

(34) Wold, S.; Sjöström, M.; Eriksson, L. PLS Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(35) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(36) Jackson, J. E. *A User's Guide to Principal Components*; Wiley: New York, 1991.

(37) Nelson, P. R. C.; Taylor, P. A.; MacGregor, J. F. Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations. *Chemom. Intell. Lab. Syst.* **1996**, *35*, 45−65.

(38) Little, R. J. A.; Rubin, D. B. *Statistical Analysis with Missing Data*, 2nd ed.; Wiley: New York, 2002.
(39) Schneider, T.; Neumaier, A. Algorithm 808: ARFIT—A Matlab Package for the Estimation of Parameters and Eigenmodes of Multivariate Autoregressive Models. *ACM Trans. Math. Software* **2001**, *27* (1), 58−65.
(40) Wood, R. K.; Berry, M. W. Terminal Composition Control of a Binary Distillation Column. *Chem. Eng. Sci.* **1973**, *28* (9), 1707−1717.
(41) Lakshminarayanan, S.; Shah, S. L.; Nandakumar, K. Modeling and Control of Multivariable Processes: Dynamic PLS Approach. *AIChE J.* **1997**, *43* (9), 2307−2322.
(42) Wachs, A.; Lewin, D. R. Improved PCA Methods for Process Disturbance and Failure Identification. *AIChE J.* **1999**, *45* (8), 1688−1700.
(43) Lin, W.; Qian, Y.; Li, X. Nonlinear Dynamic Principal Component Analysis for on-Line Process Monitoring and Diagnosis. *Comput. Chem. Eng.* **2000**, *24*, 423−429.
(44) Negiz, A.; Çinar, A. PLS, Balanced, and Canonical Variate Realization Techniques for Identifying VARMA Models in State Space. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 209−221.
(45) Shi, R.; MacGregor, J. F. Modeling of Dynamic Systems Using Latent Variable and Subspace Methods. *J. Chemom.* **2000**, *14*, 423−439.
(46) Overschee, P. V.; De Moor, B. A Unifying Theorem for Three Subspace System Identification Algorithms. *Automatica* **1995**, *31* (12), 1853−1864.
(47) Ricker, N. L. The Use of Biased Least-Squares Estimators for Parameters in Discrete-Time Pulse-Response Models. *Ind. Eng. Chem. Res.* **1988**, *27*, 343−350.
(48) Sharmin, R.; Sundararaj, U.; Shah, S.; Griend, L. V.; Sun, Y.-J. Inferential Sensors for Estimation of Polymer Quality Parameters: Industrial Application of a PLS-Based Soft-Sensor for a LDPE Plant. *Chem. Eng. Sci.* **2006**, *61*, 6372−6384.
(49) Arteaga, F.; Ferrer, A. Dealing with Missing Data in MSPC: Several Methods, Different Interpretations, Some Examples. *J. Chemom.* **2002**, *16*, 408−418.
(50) Tracy, N. D.; Young, J. C.; Mason, R. L. Multivariate Control Charts for Individual Observations. *J. Qual. Technol.* **1992**, *24* (2), 88−95.
(51) Qin, S. J. Statistical Process Monitoring: Basics and beyond. *J. Chemom.* **2003**, *17*, 480−502.
(52) Cinar, A.; Palazoglu, A.; Kayihan, F. *Chemical Process Performance Evaluation*; CRC Press: Boca Raton, FL, 2007.
(53) Nomikos, P.; MacGregor, J. F. Monitoring Batch Processes Using Multiway Principal Component Analysis. *AIChE J.* **1994**, *40* (8), 1361−1375.