Chemometrics and Intelligent Laboratory Systems 167 (2017) 1–11

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

On the structure of dynamic principal component analysis used in statistical process monitoring

Erik Vanhatalo [a, *], Murat Kulahci [b, a], Bjarne Bergquist [a]

[a] Luleå University of Technology, SE-97187 Luleå, Sweden
[b] Technical University of Denmark, Richard Petersens Plads, 2800 Kongens Lyngby, Denmark

ARTICLE INFO

Keywords:
Dynamic principal component analysis
Vector autoregressive process
Vector moving average process
Autocorrelation
Simulation
Tennessee Eastman process simulator

ABSTRACT

When principal component analysis (PCA) is used for statistical process monitoring it relies on the assumption that data are time independent. However, industrial data will often exhibit serial correlation. Dynamic PCA (DPCA) has been suggested as a remedy for high-dimensional and time-dependent data. In DPCA the input matrix is augmented by adding time-lagged values of the variables. In building a DPCA model the analyst needs to decide on (1) the number of lags to add, and (2) given a specific lag structure, how many principal components to retain. In this article we propose a new analyst driven method to determine the maximum number of lags in DPCA with a foundation in multivariate time series analysis. The method is based on the behavior of the eigenvalues of the lagged autocorrelation and partial autocorrelation matrices. Given a specific lag structure we also propose a method for determining the number of principal components to retain. The number of retained principal components is determined by visual inspection of the serial correlation in the squared prediction error statistic, Q (SPE), together with the cumulative explained variance of the model. The methods are illustrated using simulated vector autoregressive and moving average data, and tested on Tennessee Eastman process data.

1. Introduction

The need for latent variable methods in statistical process monitoring (SPM) such as principal component analysis (PCA) has grown alongside the development of automatic data collection schemes that generate multivariate data. When PCA is used for inferential purposes, such as in SPM, it relies on the assumption that data are independent in time, see Jolliffe [1] and Vanhatalo and Kulahci [2]. In today's applications, however, this assumption is becoming increasingly unrealistic. Due to system dynamics combined with high-frequency sampling, successive observations will often be serially correlated, see Montgomery et al. [3] and Bisgaard and Kulahci [4]. The negative impact of autocorrelation on PCA-based SPM has also been demonstrated by Vanhatalo and Kulahci [2].

A suggested remedy, put forward predominantly within chemometrics literature, is to augment the input and/or output matrix by adding time-lagged values of the variables and thereby expanding the data matrix. PCA performed on this expanded data matrix was introduced as Dynamic PCA (DPCA) by Ku et al. [5], see also Kourti and MacGregor [6]. DPCA implicitly assumes that data follow an autoregressive structure. By adding lagged versions of the original variables the idea is to 'convert' autocorrelation in the original variables into cross-correlation which can be modeled by PCA. DPCA is designed to account for the dynamic structure in the original data. However, the resulting principal components (scores) from DPCA may still be autocorrelated, and possibly cross-correlated, see De Ketelaere et al. [7]. These principal components (PCs) require some procedure, such as ARMA filters, to remove the autocorrelation and improve the performance in SPM. Rato and Reis [8] suggested the DPCA-DR procedure to decorrelate the monitoring statistics.

An innate issue in DPCA is to determine the number of lags of the original variables to be added to properly capture system dynamics. The available methods to identify the lag structure, such as the original method in Ku et al. [5] or the algorithm proposed by Rato and Reis [9] are typically based on algorithms using several iterations of PCA or singular value decomposition of the extended matrix. These methods have the advantage of being easy to automate but from a multivariate time series standpoint, we argue that the resulting lag recommendations may lack intuitive appeal.

In this article we propose a new analyst driven method to determine the maximum number of lags in DPCA. The method is based on analyzing the eigenvalues of the lagged autocorrelation and partial autocorrelation

* Corresponding author.
E-mail addresses: erik.vanhatalo@ltu.se (E. Vanhatalo), muku@dtu.dk (M. Kulahci), bjarne.bergquist@ltu.se (B. Bergquist).

http://dx.doi.org/10.1016/j.chemolab.2017.05.016
Received 2 February 2017; Received in revised form 15 May 2017; Accepted 16 May 2017
Available online 18 May 2017

List of abbreviations:

ARMA	Autoregressive moving average
AVE	Absolute value of the eigenvalues
DPCA	Dynamic principal component analysis
KSG-95	Lag identification algorithm from Ku et al. (1995)
PCA	Principal component analysis
PC	Principal component
Q/SPE	Squared prediction error statistic
RR-13	Lag identification algorithm from Rato & Reis (2013)
SPM	Statistical process monitoring
VAR	Vector autoregressive
VMA	Vector moving average

matrices of the original data. A visual method to choose the number of PCs to retain taking into account the autocorrelation in the residual space (Q) is also provided. We explain and illustrate how the required number of lags in DPCA differs between two extremes: vector autoregressive (VAR) and vector moving average (VMA) processes. We also illustrate the potential negative impact of adding more lags than necessary. We finally provide a small case study testing our proposed method using data from the Tennessee Eastman process simulator.

2. PCA-based SPM

This section provides a short technical background on PCA with a focus on its use in SPM. PCA can be referred to as 'static PCA' in comparison with dynamic PCA, see De Ketelaere et al. [7]. For a more complete explanation of PCA see, for example, Johnson and Wichern [10], Jackson [11] or Jolliffe [1].

To reduce dimensionality PCA extracts a few, uncorrelated principal components (PCs) as linear combinations of the original variables. Let $\mathbf{X}' = [x_1, x_2, \ldots, x_m]$ represent a random vector describing an m-dimensional variable with covariance matrix $\mathbf{\Sigma}$. Let $\mathbf{\Sigma}$ have the eigenvalue–eigenvector pairs $(\lambda_1, \mathbf{p}_1)$, $(\lambda_2, \mathbf{p}_2)$, …, $(\lambda_m, \mathbf{p}_m)$. The m PCs are orthogonal to one another and ordered with respect to their variances (eigenvalues of the covariance matrix $\mathbf{\Sigma}$). The eigenvectors, $\mathbf{p}_a$, $a = 1,2, \ldots, m$, have unit length, $\mathbf{p}_a\mathbf{p}_a' = 1$, and are called PC loading vectors. Since PCA is scale-dependent the variables are often mean centered and scaled to unit variance before PCA is conducted. With standardized variables, the correlation matrix of $\mathbf{X}$ is instead used to derive the eigenvector–eigenvalue pairs. In practice the covariance (correlation) matrix is unknown and needs to be estimated from the $n \times m$ data matrix $\mathbf{X}$.

The values of the PCs for each observation (the scores) are collected in the score vectors, $\mathbf{t}_a$, $a = 1,2, \ldots, m$. The loading vectors, $\mathbf{p}_a$, $a = 1,2, \ldots, A$, where $A < m$, define the reduced dimension space (A) with respect to the original variables and the score vectors, $\mathbf{t}_a$, $a = 1,2, \ldots, A$, are the projection of the original observations onto this reduced dimensional space. The term 'PCA model' is used to denote the approximation of the variability in $\mathbf{X}$ by the use of the A first PCs. We can write:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' = \sum_{a=1}^{A} \mathbf{t}_a\mathbf{P}_a' + \mathbf{E} \tag{1}$$

where the remaining PCs, expected to explain noise, are summed up in a matrix of residuals $\mathbf{E}$.

First of all, a subset of process data when the process is running normally is collected and the in-control PCA model is built. This is called the Phase I sample, which is also used for setting the statistically based limits in the control charts. The real-time monitoring phase in SPM is called Phase II.

Two monitoring statistics are typically used in PCA-based SPM. The scores of the A retained PCs are typically monitored in a Hotelling T^2 chart. Another common statistic is the sum of squared row-elements of the error matrix called the squared prediction error (SPE) or Q statistic. An unusual value in the T^2 chart can be viewed as an outlier within the model, i.e., in the dimensions of the retained PCs. Such an observation could indicate unusual values in some (or all) of the original variables, while the correlation structure among the variables remains more or less intact. An outlier in the Q chart can be viewed as an outlier outside the model, that is, in the dimensions of the PCs not included in the model. Such an outlier indicates that the correlation structure among the variables captured in Phase I is different from that being observed in the observation(s) in Phase II.

The Phase II upper control limit for the T^2 statistic is based on the F distribution and given as:

$$\mathrm{UCL}_{T^2} = \frac{A(n+1)(n-1)}{n^2 - nA} F_{\alpha,A,n-A} \tag{2}$$

where A is the number of retained principal components, n is the number of samples (i.e. observations) in Phase I, α is the acceptable false alarm rate, and $F_{\alpha,A,n-A}$ is the upper α percentile of the F distribution with A and n-A degrees of freedom. The Phase I upper control limit (not used in this article) for the T^2 statistic is based on the beta distribution and provided by Tracy et al. [12]. An approximate upper control limit for the Q statistics is provided by Jackson and Mudholkar [13]:

$$\mathrm{UCL}_Q = \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{3}$$

where z_α is the 100(1-α) percentile of the standardized normal distribution, $\theta_i = \sum_{j=A+1}^{p} \lambda_j^i$, and $h_0 = 1 - (2\theta_1\theta_3)/3\theta_2^2$. We note, however, that the theoretical control limits above assume time-independent data.

2.1. Dynamic PCA

The monitored processes often behave dynamically and the observations on the variables are serially dependent (autocorrelated). Ku et al. [5] introduced dynamic PCA (DPCA) to extend PCA to account for dynamic process behavior and improve monitoring performance. DPCA works in the same way as PCA but uses a 'time lag shift' method to transform the original $\mathbf{X}$ matrix into a new expanded matrix, $\tilde{\mathbf{X}}$, with l time-shifted duplicate vectors of all the variables in $\mathbf{X}$. PCA is then performed on this $\tilde{\mathbf{X}}$ matrix.

By a proper choice of l and the proper choice of retained PCs Ku et al. [5] claim that linear constraints should appear in the noise subspace (or approximate null space) with small eigenvalues. Therefore Ku et al. [5] argue that the Q statistic (sum of squared errors, SPE, from the DPCA model) will be independent and provide a means to monitor the dynamic system. Ku et al. [5] state that l is usually 1 or 2 but could be higher for nonlinear systems.

Note that DPCA implies that the resulting PCs will form an implicit vector autoregressive (VAR) model of the data. This structure can be illustrated by a simple case with two variables $x_{1,t}$ and $x_{2,t}$, where the two PCs are given as

$$\begin{aligned} PC_{1,t} &= \mathbf{P}_1'\mathbf{X}_t = p_{11}x_{1,t} + p_{12}x_{2,t} \\ PC_{2,t} &= \mathbf{P}_2'\mathbf{X}_t = p_{21}x_{1,t} + p_{22}x_{2,t} \end{aligned} \tag{4}$$

where e.g. $\mathbf{p}_1$ vector contains the loadings for PC 1. Now assume that two variables exhibit autocorrelation and that we add variables with one lag. This will produce $\tilde{\mathbf{X}}$ including the four variables, $x_{1,t}$, $x_{2,t}$, $x_{1,t-1}$, $x_{2,t-1}$ (original and lagged) and after applying DPCA the four PCs can be written as

$$\begin{aligned} PC_{1,t} &= \mathbf{P}_1'\tilde{\mathbf{X}}_{t:t-1} = p_{11}x_{1,t} + p_{12}x_{2,t} + p_{13}x_{1,t-1} + p_{14}x_{2,t-1} \\ PC_{2,t} &= \mathbf{P}_2'\tilde{\mathbf{X}}_{t:t-1} = p_{21}x_{1,t} + p_{22}x_{2,t} + p_{23}x_{1,t-1} + p_{24}x_{2,t-1} \\ PC_{3,t} &= \mathbf{P}_3'\tilde{\mathbf{X}}_{t:t-1} = p_{31}x_{1,t} + p_{32}x_{2,t} + p_{33}x_{1,t-1} + p_{34}x_{2,t-1} \\ PC_{4,t} &= \mathbf{P}_4'\tilde{\mathbf{X}}_{t:t-1} = p_{41}x_{1,t} + p_{42}x_{2,t} + p_{43}x_{1,t-1} + p_{44}x_{2,t-1} \end{aligned} \quad (5)$$

This means that, e.g. the first principal component in (5) can be described as being 'regressed on' $x_{1,t}$ and $x_{1,t-1}$ but also on $x_{2,t}$ and $x_{2,t-1}$. As Ku et al. [5] states this also means that *"processes with moving average (MA) terms will be approximated by an AR model."*

2.2. Lags in DPCA

We provide summaries of two well-described benchmark methods from the literature that analyze the eigenvalues of the expanded matrix to determine the number of lags in DPCA. The first method, here called KSG-95, is given in the original paper by Ku et al. [5] and the second method, here called RR-13, is a suggested improvement to estimate a single lag structure for all variables by Rato and Reis [9]. We note that Rato and Reis [9] also introduce an algorithm to determine individual and different lags for each variable in the original $\mathbf{X}$ matrix. However, here we focus on methods that add the same number of lags for each variable. Pseudo-codes for the two methods are provided in Table 1.

Both methods in Table 1 provide the maximum number of lags, l, to be added. The KSG-95 method identifies the lag structure by trying to determine if the PCs of the expanded matrix, $\tilde{\mathbf{X}}$, represent linear relationships followed by analysis of the autocorrelation and cross correlation plots of the scores. The RR-13 method on the other hand tries simultaneously to minimize, by choice of l, the "key singular value" and the "key singular value ratio" given $\tilde{\mathbf{X}}$. Other identification methods to determine the lag structure include the method of adding lags to minimize the determinant of the correlation matrix assuming a two-block, input-output situation given in Wachs and Lewin [14]. Another example is a method based on the Akaike information criterion [15] drawn from the signal processing community.

3. Proposed new method

We argue that the maximum number of lags may be identified by studying the pattern in the eigenvalues of both the sample autocorrelation and partial autocorrelation matrices of the original data at different lags. Below we illustrate the method with two cases: a vector moving average (VMA) process and a vector autoregressive (VAR) process. We have used the free R statistics software (The R Foundation for Statistical Computing) for all simulations, plots and methods in this article and the code is available upon request from the first author.

3.1. Eigenvalues of autocorrelation and partial autocorrelation matrices for vector MA and vector AR process

The vector MA process can be expressed as

$$\mathbf{Y}_t = \boldsymbol{\mu} - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\varepsilon}_t \quad (6)$$

where $\boldsymbol{\Theta}_j$ is the jth order MA coefficient matrix and $\boldsymbol{\varepsilon}_t$ are normally distributed errors with covariance matrix $\boldsymbol{\Sigma}$. Reinsel [16] provides the covariance matrix of a vector MA(q) model at lag k as

$$\boldsymbol{\Gamma}(k) = \sum_{h=0}^{q-k} \boldsymbol{\Theta}_h \boldsymbol{\Sigma} \boldsymbol{\Theta}_{h+k}' \quad (7)$$

where q is the order of the vector MA model, $\boldsymbol{\Theta}_0 = -I$, $\boldsymbol{\Gamma}(-k) = \boldsymbol{\Gamma}(k)'$, and $\boldsymbol{\Gamma}(k) = 0$ for $k > q$. This means that all cross-correlations are zero for lags greater than q for the vector MA(q) process.

The vector AR process can be expressed as

$$\mathbf{Y}_t = \boldsymbol{\delta} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \mathbf{Y}_{t-j} + \boldsymbol{\varepsilon}_t \quad (8)$$

where p is the order of the vector AR model, $\boldsymbol{\Phi}_j$ is the jth order AR coefficient matrix and $\boldsymbol{\varepsilon}_t$ are normally distributed errors with covariance matrix $\boldsymbol{\Sigma}$. Montgomery et al. [17] give the covariance matrix of a stationary vector AR(p) model at lag k as

$$\begin{aligned} &\boldsymbol{\Gamma}(k) = \sum_{h=0}^{p} \boldsymbol{\Gamma}(k-h)\boldsymbol{\Phi}_h' \\ &\text{and} \\ &\boldsymbol{\Gamma}(0) = \sum_{h=0}^{p} \boldsymbol{\Gamma}(-h)\boldsymbol{\Phi}_h' + \boldsymbol{\Sigma} \end{aligned} \quad (9)$$

Since we use the first order vector AR model, VAR(1), more extensively in what follows, we here provide some more details from Reinsel [16] on its covariance features

$$\begin{aligned} &\boldsymbol{\Gamma}(0) = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(0) \boldsymbol{\Phi}_1' + \boldsymbol{\Sigma}, \\ &\text{and} \\ &\boldsymbol{\Gamma}(k) = \boldsymbol{\Gamma}(0) \boldsymbol{\Phi}_1'^{k}, \quad k \geq 1 \end{aligned} \quad (10)$$

Furthermore, the correlation matrices $\boldsymbol{\rho}(k)$ have the form

$$\begin{aligned} \boldsymbol{\rho}(k) &= \boldsymbol{\rho}(0)\mathbf{V}^{-1/2}\boldsymbol{\Gamma}(k)\mathbf{V}^{-1/2} \\ &= \boldsymbol{\rho}(0)\mathbf{V}^{-1/2}\boldsymbol{\Phi}_1'^{k}\mathbf{V}^{-1/2} \end{aligned} \quad (11)$$

where $\mathbf{V} = diag(\gamma_{11}(0), \gamma_{22}(0), \ldots, \gamma_{mm}(0))$, that is a diagonal matrix with the auto-covariance coefficients of the individual time series, $\gamma_{ii}(0)$, on the diagonal.

Table 1
Pseudo-code for the KSG-95 and RR-13 methods.

KSG-95 (Ku et al., 1995)	RR-13 (Rato and Reis, 2013)
1. Set $l = 0$.	1. Set $l = 0$.
2. Form new data matrix, $\tilde{\mathbf{X}} = [\mathbf{X}(k)\ldots\mathbf{X}(k-l)]$.	2. Form new data matrix, $\tilde{\mathbf{X}} = [\mathbf{X}(k)\ldots\mathbf{X}(k-l)]$.
3. Perform PCA on $\tilde{\mathbf{X}}$ and all principal component scores.	3. Perform the singular value decomposition of the covariance of the extended matrix, $\Sigma_{\tilde{\mathbf{X}}} = USV^T$.
4. Set $h = m \times (l+1)$ and $r(l) = 0$.	4. Set $KSV(l) = s_{ml+1}$.
5. Determine if the hth component represents a linear relation (check of matrix rank deficiency). If yes, proceed, if no go to step 7.	5. If $l > 0$ set $KSVR(l) = KSV(l)/KSV(l-1)$.
6. Set $h = h-1$ and $r = r(l)+1$, repeat step 5.	6. If $l < l_{\max}$ set $l = l+1$ and go to step 2.
7. Calculate the number of new relationships: $r_{new}(l) = r(l) - \sum_{i=0}^{l-1}(l-i+1)r_{new}(i)$	7. Normalize KSV and $KSVR$.
8. If $r_{new}(l) \leq 0$, go to step 10, else proceed.	8. Determine $\text{argmin}_{l\in[1,l_{\max}]}\sqrt{KSV_N(l)^2 + KSVR_N(l)^2}$, s.t. $l \geq l^*$ and l^* is the first l such that $KSVR(l) < KSVR(l-1)$.
9. Set $l = l + 1$, go to step 2.	
10. Stop.	
Notes: m = number of variables r = rank or number of relations	Notes: s_{ml+1} is the $(ml+1)^{th}$ singular value of $\Sigma_{\tilde{\mathbf{X}}}$. KSV = "Key singular value", KSVR = "Key singular value ratio".

Reinsel [16] also defines the partial cross-correlation matrix at lag k, P_k, for vector processes which we here denote the partial autocorrelation matrix at lag k. The matrix P_k is the cross-correlation matrix between the elements of $\mathbf{Y}_{t-k}$ and $\mathbf{Y}_t$, after adjusting them both for their dependence on the intervening values $\mathbf{Y}_{t-k+1}, \ldots, \mathbf{Y}_{t-1}$. The (i, j)th element of P_k is the partial correlation between variables $Y_{i,t-k}$ and $Y_{j,t}$ adjusting for $\mathbf{Y}_{t-k+1}, \ldots, \mathbf{Y}_{t-1}$. For a VAR($p$) process P_k has the 'cutoff' property such that $P_k = 0$ for all $k > p$. The reader is referred to Reinsel [16] for further details.

Our proposed method uses the pattern in the magnitudes of the eigenvalues of the sample correlation matrices, $\boldsymbol{\rho}(k)$. It also uses the sample partial autocorrelation matrices, P_k, for different time lags, k, to identify the type and the order of vector process and consequently the maximum lags required in DPCA. We note that none of the eigenvalues will be exactly zero for sample data. However, when all eigenvalues are small (close to zero) it is an indication of little or no remaining 'information' in the lagged autocorrelation or partial autocorrelation matrix for the specific lag. In other words, if all eigenvalues of sample autocorrelation and partial autocorrelation matrices of lag k are small, it can be assumed that no linear relationship exists between $\mathbf{X}_t$ and $\mathbf{X}_{t-k}$ for any of the variables in $\mathbf{X}_t$ and hence there is no need to further expand $\tilde{\mathbf{X}}$ by adding $\mathbf{X}_{t-k}$.

Note that a process with MA terms needs to be approximated with an implicit VAR model in DPCA. We therefore argue that the identified order of the VAR process should be the same as the maximum number of lags required in DPCA. As a consequence we expect MA type data to require more lags than AR type data. Note that the interpretations of our method assume that we are dealing with stationary VAR and invertible VMA processes. If the VMA process is invertible it is possible to approximate it with a high order VAR model. As discussed in greater detail by Reinsel [16], the VMA(q) process is invertible if all roots of $\det\{I - \boldsymbol{\Theta}_1 B - \ldots - \boldsymbol{\Theta}_q B^q\} = 0$ are greater than one in absolute value. Similarly, the VAR(p) process is stationary if all roots of $\det\{I - \boldsymbol{\Phi}_1 B - \ldots - \boldsymbol{\Phi}_p B^p\} = 0$ are greater than one in absolute value.

We plot the absolute value of the eigenvalues (AVE) of the sample lagged autocorrelation and partial autocorrelation matrices of the original data matrix to identify the maximum number of lags in DPCA. The AVE (or equivalently, the length of the eigenvalue, sometimes also known as the modulus) is used since some eigenvalues may be complex numbers. A summary of the important patterns in these plots exhibited by VMA and VAR processes are given in Table 2. Compared to the RR-13 method our proposed method has the advantage that it does not assume that at least 1 lag should be added. That is, in the (perhaps rare) event the analyst intends to use DPCA on time-independent data our method will show that indeed no lag should be added.

3.2. Determining the number of PCs

An important step in PCA-based SPM in Phase I, and the secondary step in DPCA after the number of lags (l) to add have been identified, is to determine the number of PCs, A, to retain in the model to capture normal process behavior and this is often difficult with noisy data. There are several ways to determine A, such as achieving a specific limit of cumulative explained variance, looking for a 'knee' in the scree plot, limits for the minimum eigenvalue and cross-validation. Ku et al. [5] argue in favor of using parallel analysis for choosing A, but at the same time recommend using several approaches in combination to come up with the final choice.

In trying to best fulfill the goal described by Ku et al. [5] to achieve Q statistics that are independent or nearly independent in time, we argue that the proper choice of A (given that l is determined) needs to consider the autocorrelation in the Q statistics. We therefore propose to study the autocorrelation function (ACF) for the Q statistics in the Phase I sample and to choose A by trying to minimize the autocorrelation while simultaneously maintaining a high cumulative explained variance.

Table 2
Patterns in the eigenvalues of the sample lagged autocorrelation and partial autocorrelation matrices used to interpret the plots in our method.

Plots based on sample data	VMA(q) process	VAR(p) process
AVE of the autocorrelation matrices at lag k	Eigenvalues will cut off (approach zero) after lag $l > q$.	Eigenvalues will die out slowly exhibiting lingering high eigenvalues.
AVE of the partial autocorrelation matrices at lag k	Eigenvalues will die out slowly exhibiting lingering high eigenvalues.	Eigenvalues will cut off (approach zero) after lag $l > p$.

3.3. Summary of proposed method

We note that our proposed method is not an algorithm easily automated but rather a recommended procedure for the analyst largely based on visual inspection of the patterns in eigenvalues and in the autocorrelation in the Q statistic. In our opinion any method including those in Table 1 can at best provide an approximation for the "true" number of lags to be added to capture the serial dependence in data. We would recommend that relevant process knowledge of which the quantitative analyst is usually bereft is used in the analysis. Table 3 provides the important main steps of our proposed method.

4. Lags and number of PCs in DPCA for simulated vector MA and vector AR processes

In this section we explore DPCA performance using simulated time-independent data and data from VAR(1) and VMA(1) processes. For all these cases we use a five-variable model, to limit complexity, and the covariance matrix $\boldsymbol{\Sigma}$ for the errors is given as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & .9 & .8 & 0 & 0 \\ .9 & 1 & .7 & 0 & 0 \\ .8 & .7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & .9 \\ 0 & 0 & 0 & .9 & 1 \end{bmatrix} \tag{12}$$

The singular values of $\boldsymbol{\Sigma}$ are 2.603, 1.900, 0.314, 0.100, and 0.083, which essentially means that the resulting variables, for time-independent data, are correlated through two blocks of correlated errors for (x_1, x_2, x_3) and for (x_4, x_5) respectively. Based on these singular values the first two PCs explain 90.05% of the variability for time-independent data in PCA.

4.1. Time-independent data

It could be deemed irrelevant to test DPCA for time-independent data, but we provide this example for comparison and illustration purposes. Fig. 1a–b illustrate plots of the AVE of the autocorrelation and partial

Table 3
Step-by-step summary of the proposed methodology.

Step 1	Given the $n \times m$ data matrix $\mathbf{X}$, estimate its sample lagged autocorrelation matrices at lag k, $\widehat{\boldsymbol{\rho}}(k)$, for $k = 0, 1, 2, \ldots, k_{\max}$ and sample partial autocorrelation matrices at lag k, $\widehat{P}_k(k)$, for $k = 1, 2, \ldots, k_{\max}$.
Step 2	Plot the absolute value of the eigenvalues (AVE) of the sample lagged autocorrelation and partial autocorrelation matrices as a function of lag, k. Theoretical patterns from Table 2 are helpful to identify if the data mainly follow a VAR or VMA structure.
Step 3	Compare the AVEs with the maximum eigenvalue (for each k) of the lagged autocorrelation or partial autocorrelation matrices for simulated, time-independent dataset with the same correlation structure. The maximum lag $= l$ can be estimated from the lag (k) before which all eigenvalues are smaller than the maximum eigenvalue for time-independent data.
Step 4	To determine A, how many PCs to retain in the model, the sample autocorrelation function (ACF) of the Q statistics in the Phase I sample is calculated. This procedure is repeated for $A = [1; m-1]$ and A is chosen to minimize the autocorrelation in the Q statistics while simultaneously maintaining a high cumulative explained variance.

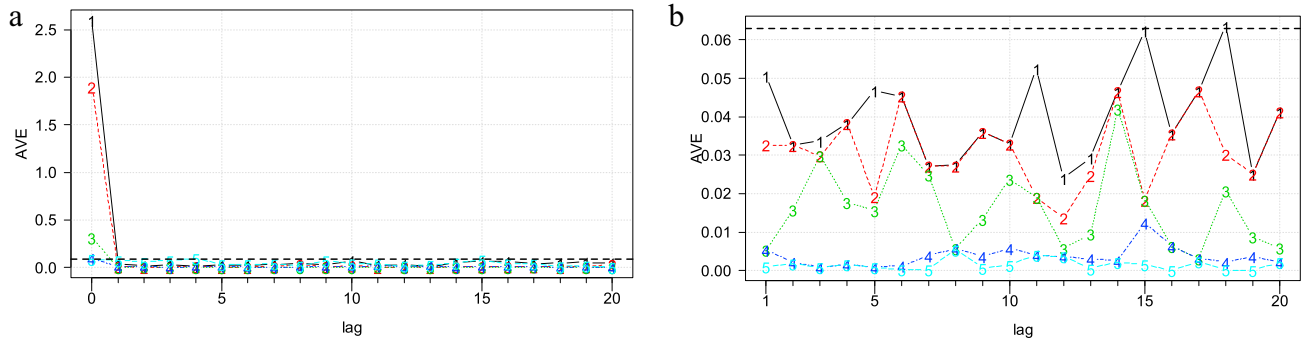

Fig. 1. **a.** AVE of the lagged autocorrelation matrices. Time-independent data. **b.** AVE of the lagged partial autocorrelation matrices. Time-independent data. Note. The horizontal dashed line represents the maximum eigenvalue for lag k of the lagged autocorrelation or partial autocorrelation matrices for another time-independent dataset with the same correlation matrix and number of observations.

autocorrelation matrices from a Phase I sample of 5000 observations of time-independent data using the covariance matrix for the errors in (12). In Fig. 1a the eigenvalues of the lagged autocorrelation matrix die out after lag 0, which is expected for time-independent data. The behavior of the eigenvalues in Fig. 1b shows no distinct pattern but seems unaffected by the lags and the eigenvalues are all small. This appearance would suggest no lag to be added in DPCA. Strictly following the KSG-95 algorithm in Table 1, excluding the more subjective recommendation of studying auto- and cross-correlation plots of the scores, the KSG-95 method suggests zero lags to be added. Applying the RR-13 method 2 lags minimize the objective function in Table 1 although 1 lag also results in nearly as small value of the objective function.

When applying static PCA to the time-independent dataset 2 PCs would explain roughly 90% of the variance in the data and the Q statistics would have low autocorrelation independent of the choice of number of PCs to retain in the model.

We move on to illustrate the impact of incorrectly using DPCA and one unnecessary lag for time-independent data. The cumulative explained variance and the absolute values of the autocorrelation function of the Q statistics as functions of the number of retained PCs in DPCA with 1 lag are presented in Fig. 2a–b. It is perhaps not surprising that incorrect use of DPCA for time-independent data results in the need of 4 retained PCs to maintain the cumulative explained variance at around 90%. It should be noted, however, that DPCA used on time-independent data will introduce substantially higher and fluctuating lag 1 autocorrelation in the Q statistics, see Fig. 2b. Using DPCA with 1 lag for time-independent data introduces positive autocorrelation in both the T^2 and Q statistics. In other tests not shown here, we can see that both the size of the autocorrelation and the number of lags with significant values in the autocorrelation functions increase with increasing number of unnecessary lags added to the original data. This in turn renders the standard deviation and thus makes the theoretical control limits in (2) and (3) underestimated.

4.2. Data from a VAR(1) process

Industrial data can often be approximated by a first-order autoregressive model. In this example we simulate VAR(1) data using a five-variable process. We continue to use the same covariance matrix for the errors, $\mathbf{\Sigma}$, as in (12). AR autocorrelation coefficient values for the $\mathbf{\Phi}$ matrix are randomly drawn from a uniform distribution between 0.3 and 0.95 to form a VAR(1) process with positively autocorrelated data. The $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ matrices used in this example are

$$\mathbf{\Phi} = \begin{bmatrix} .88 & 0 & 0 & 0 & 0 \\ 0 & .58 & 0 & 0 & 0 \\ 0 & 0 & 0.89 & 0 & 0 \\ 0 & 0 & 0 & 0.66 & 0 \\ 0 & 0 & 0 & 0 & 0.30 \end{bmatrix}; \mathbf{\Sigma} = \begin{bmatrix} 1 & .9 & .8 & 0 & 0 \\ .9 & 1 & .7 & 0 & 0 \\ .8 & .7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & .9 \\ 0 & 0 & 0 & .9 & 1 \end{bmatrix} \quad (13)$$

which produces a stationary VAR(1) process.

Fig. 3a shows the AVE of the lagged autocorrelation matrices from a Phase I sample of 5000 observations from the VAR(1) model using the $\mathbf{\Phi}$ matrix and error covariance matrix in (13). Fig. 3b shows the AVE of the partial autocorrelation matrices under the same condition.

In Fig. 3a the eigenvalues of the lagged autocorrelation matrix die out slowly while the eigenvalues in Fig. 3b show a distinct cut-off after the first lag. This appearance suggests that 1 lag should be enough in DPCA, which is expected since the linear relationships in the VAR(1) process are limited to x_t and x_{t-1}.

Strictly following the KSG-95 algorithm in Table 1, excluding the more subjective recommendation of studying auto- and cross-correlation plots of the scores, the KSG-95 method suggests adding one lag. In the RR-13 method 2 lags minimize the objective function in Table 1 but both 1 and 3 lags result in small values of the objective function.

Given a choice to add one lag in DPCA, the next step is selecting the number of PCs to retain in the DPCA model. Here we evaluate Fig. 4a–b

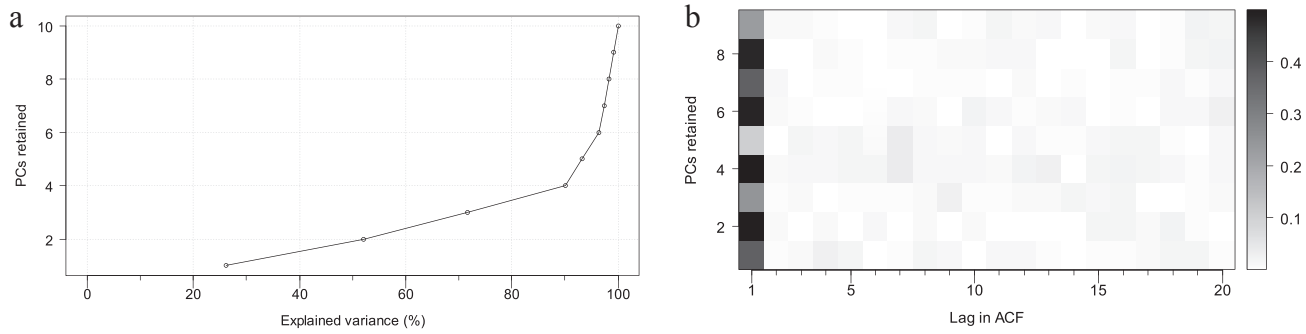

Fig. 2. **a.** Cumulative explained variance given the number of retained PCs in DPCA with 1 lag. **2b.** Absolute values of the autocorrelation function for the Q statistics given the number of retained PCs in DPCA with 1 lag.

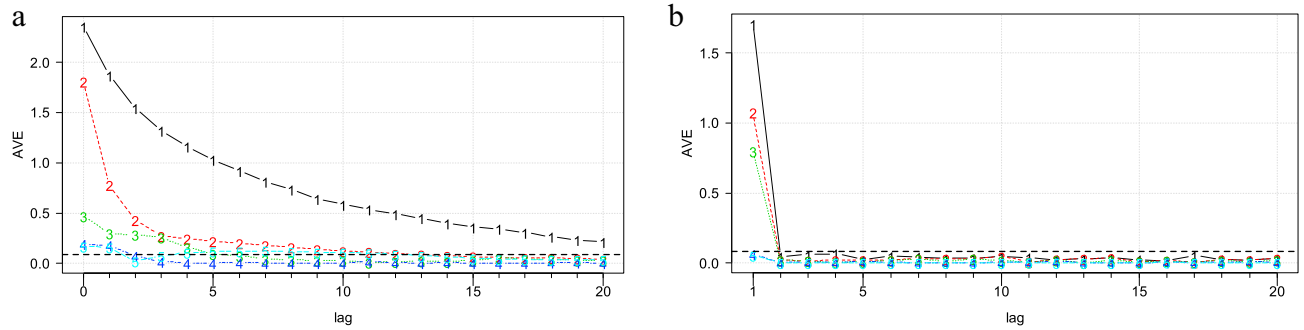

Fig. 3. **a.** AVE of the lagged autocorrelation matrices. VAR(1) data. **3b.** AVE of the lagged partial autocorrelation matrices. VAR(1) data. Note. The horizontal dashed line represents the maximum eigenvalue for lag k of the lagged autocorrelation or partial autocorrelation matrices for another time-independent dataset with the same correlation matrix and number of observations as the simulated VAR(1) data.

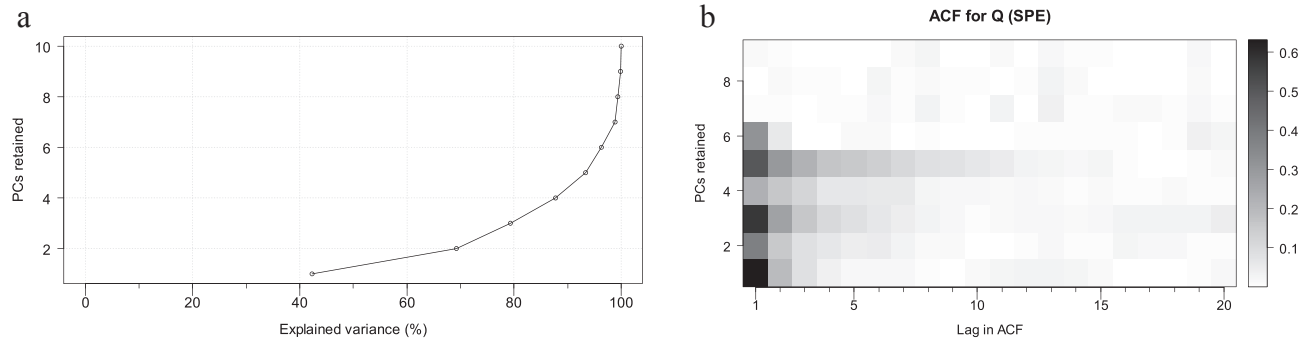

Fig. 4. **a.** Cumulative explained variance given the number of retained PCs in DPCA with 1 lag. **4b.** Absolute values of the autocorrelation function for the Q statistics given the number of retained PCs in DPCA with 1 lag.

with the cumulative explained variance and the absolute values of the autocorrelation function of the Q statistics as functions of the number of retained PCs in DPCA with 1 lag.

We try to minimize the autocorrelation in the Q statistic in Phase I while at the same time considering the cumulative explained variance to select the number of PCs to retain. From Fig. 4b we can conclude that four retained PCs would render the lag 1 autocorrelation in the Q statistic to be around 0.2. The model would then explain just below 90% of the variance. Indeed, if we retain 7 PCs or more there is very little autocorrelation left in the Q statistics but the model would explain close to 100%

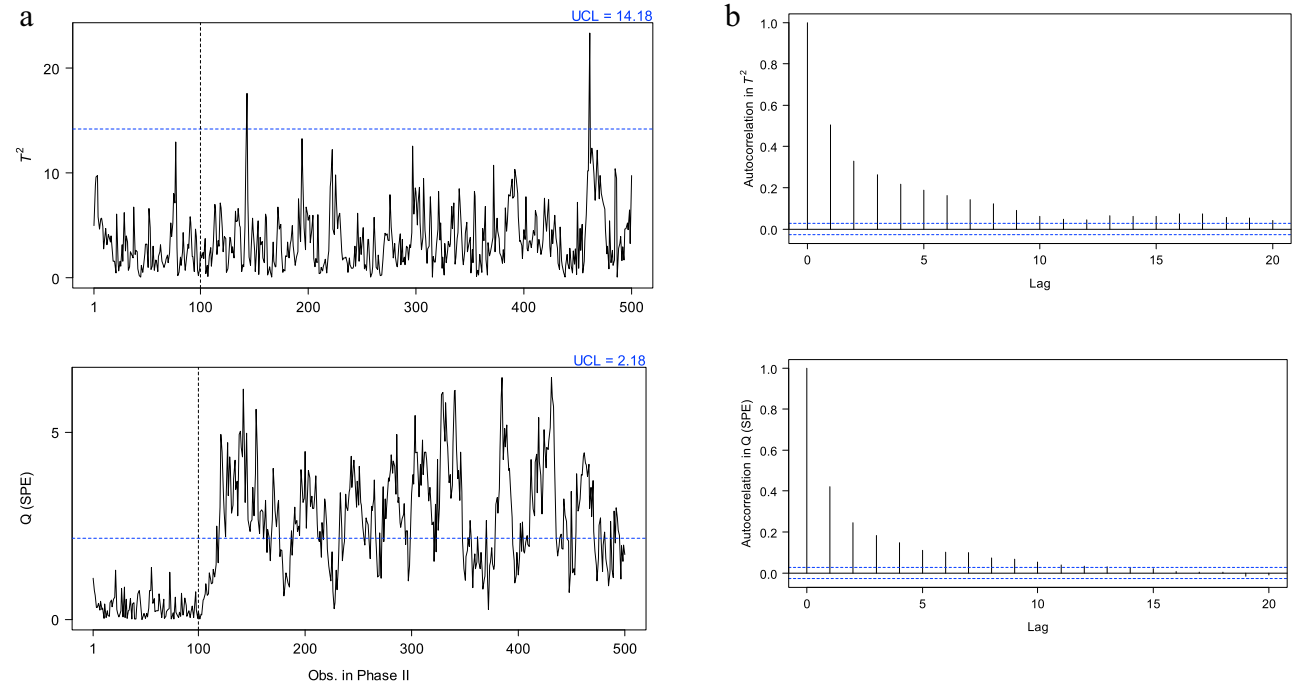

Fig. 5. **a.** Hotelling T^2 and Q charts in Phase II for Static PCA and 3 retained PCs. **5b.** Sample ACF of the T^2 and Q values in Phase I.

of the variance, which may be argued to be an overfitted model.

Next we study the monitoring performance comparing static PCA and DPCA and in these examples we use 500 observations of Phase II data where a two standard deviation step-shift in the first variable occurs at the 100th observation in the Phase II sample. We use static PCA for monitoring for comparison. In deciding on how many PCs to retain using static PCA the choice was between two or three PCs. Since three retained PCs seemed to provide the best monitoring performance for Phase II data and to provide the fairest comparison we chose three PCs in the example we illustrate in Fig. 5a–b. Figs. 5–6 provide process monitoring charts for static PCA and DPCA with one lag for 500 observations of Phase II data. The theoretical upper control limits in Figs. 5–6 are based on a 99.73% confidence level. It should however be noted that theoretical limits are not well adjusted for autocorrelated data, which especially affects the Hotelling T^2 chart. We are currently working on addressing this issue in our ongoing research.

Comparing Figs. 5 and 6 we conclude that in a one lag DPCA model with four retained PCs, the autocorrelation in the Q chart is reduced somewhat from 0.4 to around 0.2 for the lag 1 autocorrelation in the Q statistics. We also tested to retain as many as seven PCs (model not shown here) but then the signal is transferred from the residual dimensions to the model dimensions and instead shows in the T^2 chart. This phenomenon illustrates the 'traditional' interpretations of the two charts may be affected by how many PCs are retained in DPCA. The answer to the question on which chart the out of control signal will appear depends both on the characteristics of the shift as well as the number of PCs retained. Furthermore, for VAR(1) data we have not been able to achieve a model with a lag structure and number of retained PCs such that the Q statistics become time independent while at the same being able to detect the shift. In other words, we are not able to fully achieve the time independence in Q that may be expected from the work by Ku et al. [5].

The autocorrelation in the Q statistics can however be reduced substantially by choosing the number of lags in DPCA properly and by retaining a suitable number of PCs. This will make the theoretical control limit in the Q chart better suited for monitoring.

When the appropriate number of lags are added and the appropriate number of PCs are retained we expect that the autocorrelation in the Q statistics will be reduced and the false alarm rate for VAR(1) data should be reduced. The Hotelling T^2 chart will, however, suffer from high autocorrelation in the T^2 values and its performance in process monitoring will be poor. We argue that control limits may need to be adjusted for both control charts to produce reasonable Type I error rates although the autocorrelation in the Q chart can be reduced through proper use of DPCA.

4.3. Data from a VMA(1) process

DPCA implicitly models VAR data, so VMA data may perform differently using DPCA. The VMA(1) process is the simplest VMA model but is of interest to study here since the VMA(1) process has an infinite AR representation and is invertible if all eigenvalues of $\boldsymbol{\Theta}$ are less than one in absolute value, see Reinsel [16].

For this case we once again randomly pick values of the MA autocorrelation coefficients on the diagonal of $\boldsymbol{\Theta}$ from a uniform distribution between −0.3 and −0.95, producing positively autocorrelated data. The $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ matrices matrix used in this example are

$$\boldsymbol{\Theta} = \begin{bmatrix} -.74 & 0 & 0 & 0 & 0 \\ 0 & -.47 & 0 & 0 & 0 \\ 0 & 0 & -.87 & 0 & 0 \\ 0 & 0 & 0 & -.37 & 0 \\ 0 & 0 & 0 & 0 & -.39 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 1 & .9 & .8 & 0 & 0 \\ .9 & 1 & .7 & 0 & 0 \\ .8 & .7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & .9 \\ 0 & 0 & 0 & .9 & 1 \end{bmatrix} \tag{14}$$

Fig. 7a–b illustrate plots of the AVE of the lagged autocorrelation and lagged partial autocorrelation matrices from a Phase I sample. The sample includes 5000 observations from an invertible VMA(1) process with the MA coefficient matrix and the covariance matrix for the errors given in (14). In Fig. 7a the eigenvalues of the lagged autocorrelation matrix die out distinctly after the first lag while the eigenvalues in Fig. 7b shrink gradually. This is the expected behavior of VMA(1) data. The VMA(1) model, if invertible, can be written as an infinite VAR model and the linear relationships exist between $x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-\infty}$. Therefore, an invertible VMA(1) model can be reasonably approximated with high order VAR model and hence will likely require more than 1 lag in DPCA.

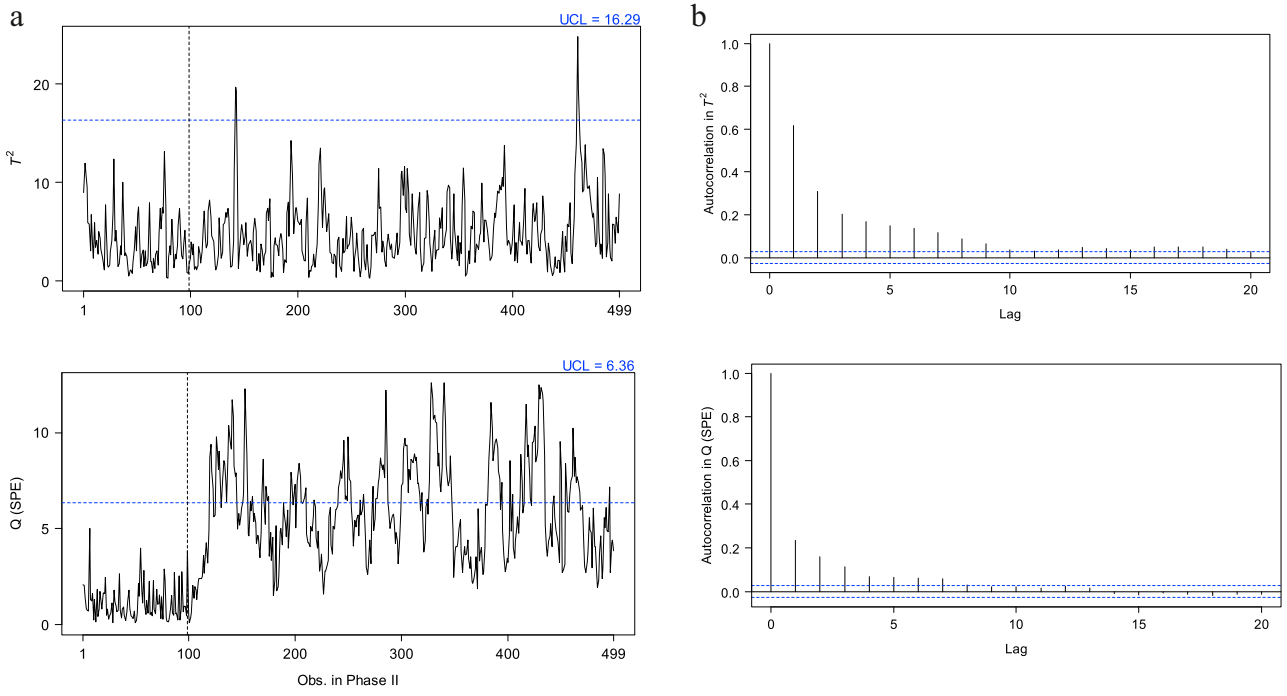

Fig. 6. a. Hotelling T^2 and Q charts for DPCA with 1 lag and 4 retained PCs. **6b.** Sample ACF of the T^2 and Q values in Phase I.

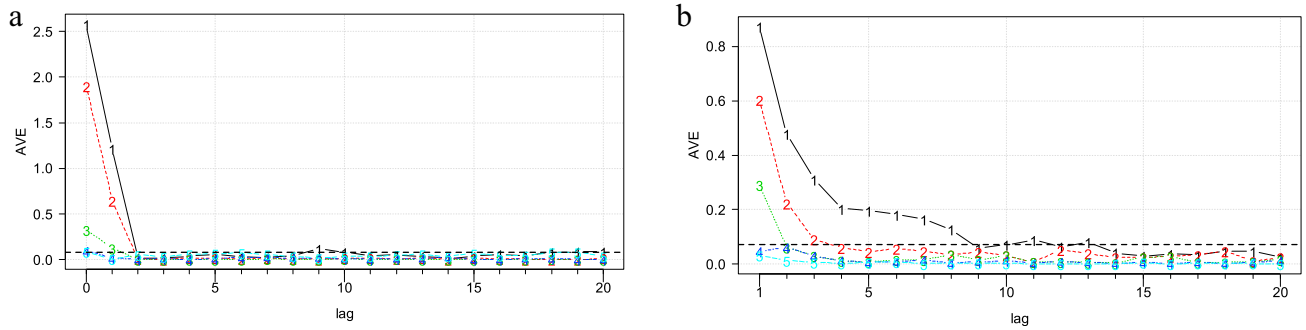

Fig. 7. a. AVE of the lagged autocorrelation matrices. VMA(1) data. **7b.** AVE of the lagged partial autocorrelation matrices. VMA(1) data.

Strictly following the KSG-95 algorithm in Table 1, excluding the more subjective recommendation of studying auto- and cross-correlation plots of the scores, the KSG-95 method suggests no lags to be added. In the RR-13 method 2 lags minimize the objective function in Table 1 but 1 lag results in nearly as small value of the objective function.

The appearance of the AVE of the lagged partial autocorrelation matrices in Fig. 7b suggest that more than one lag is necessary to approximate the VMA(1) data with DPCA properly. In fact, the visual behavior of the eigenvalues hints that as many as eight lags may be necessary although the AVEs become relatively small after three lags. Consequently, for VMA(1) data the number of lags suggested by our method and the KSG-95 and RR-13 methods differ substantially.

If we proceed with adding eight lags in DPCA, the next step will be to choose the number of PCs to retain in the DPCA model. For this, we evaluate Fig. 8a–b with the cumulative explained variance and the absolute values of the autocorrelation function of the Q statistics. As can be seen in Fig. 8a–b choosing the number of PCs to retain is a trade-off between the cumulative explained variance and the autocorrelation in the Q statistic.

To reduce the autocorrelation in the Q statistics we basically have two options: 9 PCs (~70% explained variance) or 21 PCs (>90% explained variance). To compare static PCA with two retained PCs explains 89.7% of the variance. Choosing eight lags and 21 PCs yields a lag 1 autocorrelation coefficient for the Q statistic to be around 0.6.

We proceed to test the performance of static PCA with 2 PCs (which provided the best monitoring performance in Phase II) compared with DPCA with eight lags and 21 retained PCs, see Figs. 9 and 10.

Figs. 9a and 10a are based on 500 observations of Phase II data. A two standard deviation step-shift in the first variable is generated after 100 observations in the Phase II sample. The theoretical upper control limits are based on a 99.73% confidence level. We note, however, that the theoretical limits become less appropriate as the autocorrelation in the monitoring statistics increases.

We could not find a lag structure in DPCA in combination with a number of retained PCs that substantially reduced the autocorrelation in the Q statistics for any of the examples that we have tested. The autocorrelation in the monitoring statistics for static PCA is small, below 0.2; see Fig. 9b. Adding more lags increases the autocorrelation in the T^2 and Q statistics, making the theoretical limits less appropriate for monitoring. Although the fault detection rate in the Q chart for the DPCA model with eight lags and 21 PCs is high it is partly due to the underestimated control limit. In fact, our tests with VMA(1) data indicate that static PCA is competitive due to low levels of autocorrelation in the T^2 and Q statistics. Interestingly, the KSG-95 algorithm suggested no lags to be added which means that static PCA would be recommended. In this case we also conclude that the control limits need adjustments due to the increasing autocorrelation in the plotted statistics if DPCA is used with many added lags in process monitoring.

5. Method test using Tennessee Eastman process data

In this section we apply our method to the data from the Tennessee Eastman (TE) process. The TE process has been used in previous applications of DPCA, e.g. Ku et al. [5] and Rato and Reis [9]. The TE process is a model of an industrial chemical process originally introduced by Downs and Vogel [18] for the purpose of developing process control strategies.

We use data from the TE process made available by the Braatz group [19] to compare our findings with the results reported by Rato and Reis [9]. These data are collected after implementing the plant-wide control system described in Lyman and Georgakis [20]. The TE process simulation holds 41 measurements (XMEAS) and 12 manipulated variables (XMV). There are 21 pre-programmed faults (process upsets) that can be studied. The datasets for stable operation and process upsets contain 960 observations using a sample interval of 3 min. Process faults are introduced after 8 h (or after 160 observations). Since XMV12, the agitation speed of the reactor's stirrer, has a constant value, it is excluded, resulting in 52 variables used in process monitoring.

Rato and Reis [9] compare static PCA with dynamic PCA with the

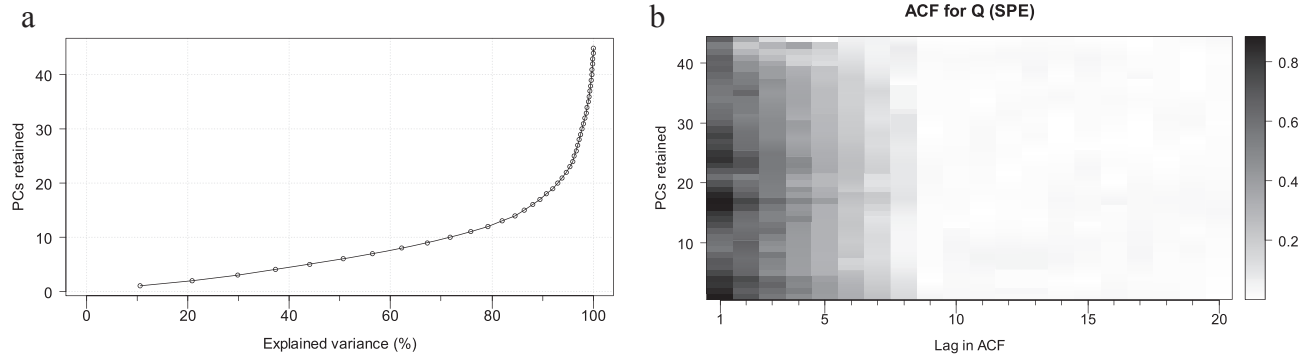

Fig. 8. a. Cumulative explained variance given the number of retained PCs in DPCA with 8 lags. **8b.** Absolute values of the autocorrelation function for the Q statistics given the number of retained PCs in DPCA with 8 lags.

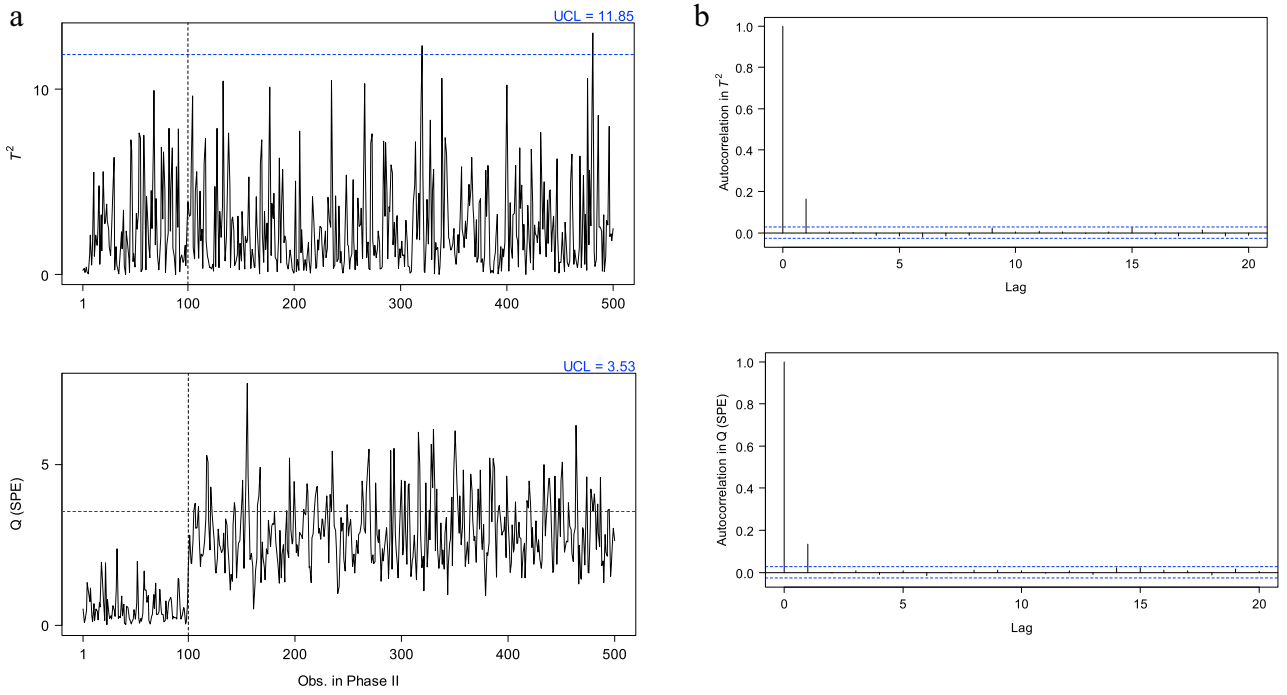

Fig. 9. **a.** Hotelling T^2 and Q charts in Phase II for Static PCA and 2 retained PCs.**b.** Sample ACF of the T^2 and Q values in Phase I.

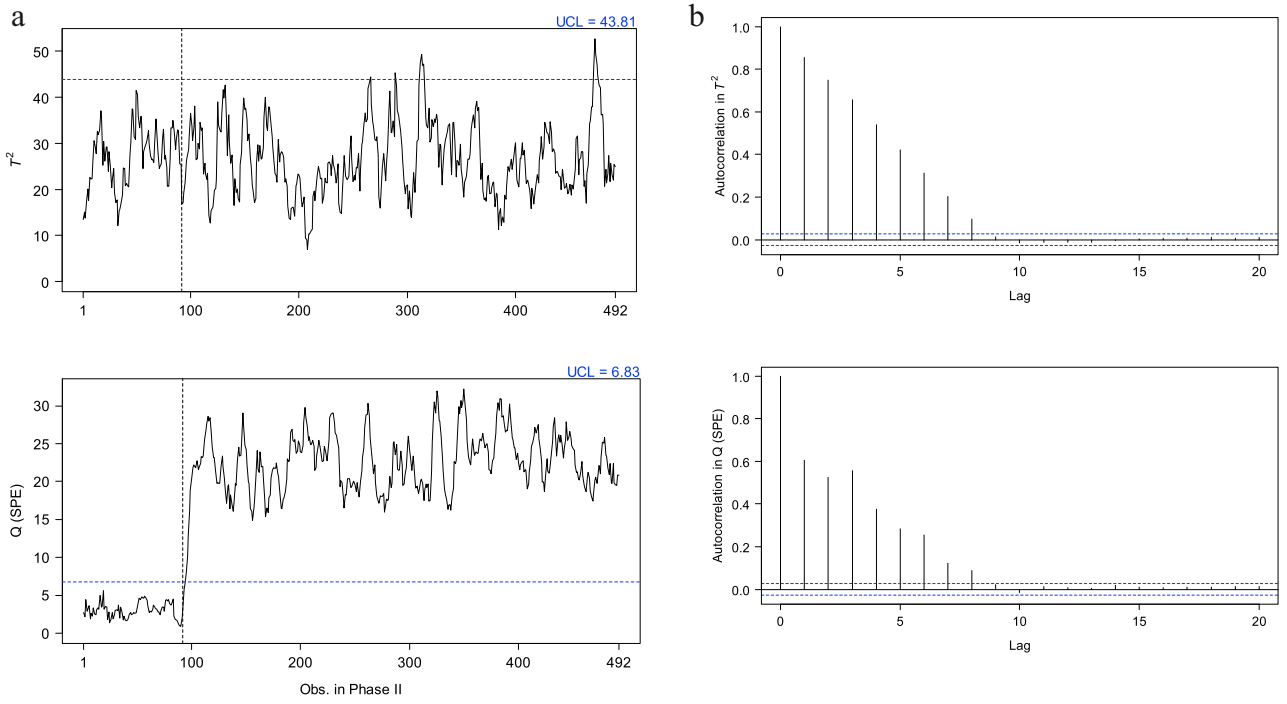

Fig. 10. **a.** Hotelling T^2 and Q charts in Phase II for DPCA with 8 lags and 21 retained PCs. **b.** Sample ACF of the T^2 and Q values in Phase I.

same number of lags for all variables. They propose a static PCA model with 17 PCs and a DPCA model (with a common lag structure for all variables) with three lags and 29 PCs.

Fig. 11a–b provides plots of the AVE of the lagged autocorrelation and partial autocorrelation matrices based on the in-control data set from the TE process.

From Fig. 11a we see that especially the first two eigenvalues die out slowly, suggesting the data have an autoregressive structure. In Fig. 11b we notice high eigenvalues until lag two or three and then the eigenvalues drop below the limit from the highest eigenvalue from a sample of time-independent data. The first eigenvalue at lag five just breaches the limit and then shrinks for higher lags. The eigenvalues increase again from lag 10 and onwards, which we argue may to be due to spurious correlations occurring due to lagging. Fig. 11b lends some support to the three lags proposed by Rato and Reis [9]. We note that in their article Rato and Reis [9] also provide an algorithm to set individual lags for each

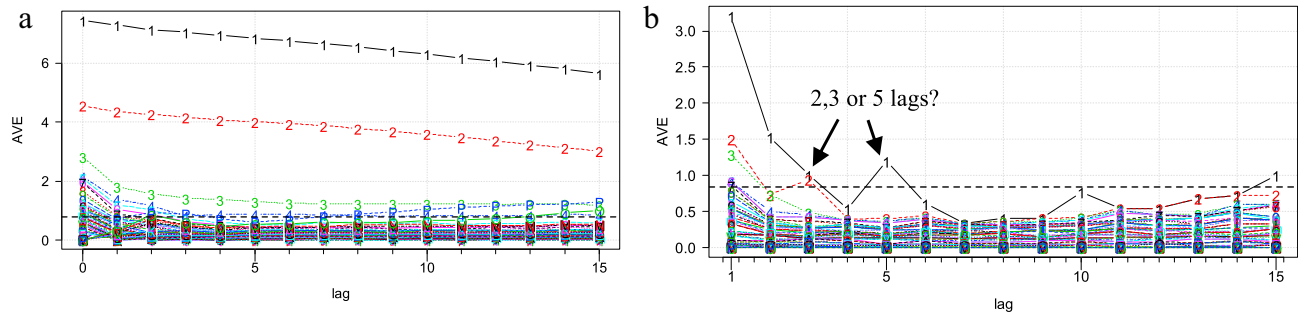

Fig. 11. **a.** AVE of the lagged autocorrelation matrices. Tennessee Eastman process data (52 variables, 960 observations). **b.** AVE of the lagged partial autocorrelation matrices. Tennessee Eastman process data (52 variables, 960 observations).

variable which gives a very different recommendation for this case with lags varying from 4 to 17 for the individual variables.

Based on the appearance of Fig. 11b we proceed under the assumption that three lags in DPCA are enough to capture process dynamics. The next step is to choose the number of PCs to retain in the DPCA model. We therefore evaluate Fig. 11a–b with the cumulative explained variance and the absolute values of the autocorrelation function of the Q statistics as functions of the number of retained PCs.

The choice is again a trade-off between the cumulative explained variance and the autocorrelation in the Q statistic. Studying Fig. 12b we can see an 'area' around 85 retained PCs where the autocorrelation in the Q statistics remain around 0.4 throughout the lags in the ACF. Retaining around 85 PCs would explain roughly 90% of the variance, see Fig. 12a. Our recommendation of the number of retained PCS differs substantially from the 29 retained PCs suggested by Rato and Reis [9]. Retaining 29 PCs would render a high lag 1 autocorrelation coefficient in the Q statistics and only around 60% explained variance. Performing parallel analysis on the TE process data with three lags would suggest retaining 38 PCs.

The TE process data are highly autocorrelated and even if we choose the number of retained PCs trying to minimize the autocorrelation, the Q statistics remains autocorrelated. Hence the theoretical control limit for the Q chart will be underestimated and the control limit should be adjusted. A remedy used in some studies, see for example Rato and Reis [9], is to split the in-control data into a calibration and validation set. The DPCA model is then built based on the calibration set and the validation set is used to adjust the limits to achieve e.g. a 1% false alarm rate for control charts. While this adjustment may improve the control limits, the adjustment is incorrect since the validation set observations are time dependent. The Type I error rate can therefore not be controlled. Russel et al. [21] describe another procedure to adjust limits of the monitored statistics. They adjust the limits to the tenth highest values of the monitored statistics for the normal operating conditions in the testing data for each of the 21 faults. While their procedure makes it possible to adjust the limits based on 21 independent data sets we expect that the adjustment will not be as good as for a large number of independent simulations in the TE process. We acknowledge that more sophisticated approaches to reduce the autocorrelation in the monitored statistics have been proposed. Kruger et al. [22] demonstrated the fitting of ARMA filters to principal component scores to remove autocorrelation. Rato and Reis [8] propose a method to reduce the autocorrelation in the monitored statistics by monitoring residuals between a one-step-ahead prediction of the scores obtained from the implicit latent variable VAR model in DPCA and the actual scores obtained in the model.

In summary our proposed visual method to determine the number of lags in DPCA show great potential in providing the analyst with a reasonable recommendation of a common lag structure for the TE process data that are in fair agreement with previous methods. Our visual method to pick the number of PCs adds additional value by providing a method to determine the number of PCs to retain to simultaneously achieve high explained variance and reduced autocorrelation in the Q statistics.

6. Conclusions and discussion

This article focuses on providing a new method of selecting the maximum number of lags in DPCA. Previous methods often study the eigenvalues of the iteratively expanded data matrix, determine its rank, and thereby provide a recommendation of how many lags to add. Our method has its foundation in multivariate time series theory and is analyst driven. We have shown that the maximum number of lags can be identified by plotting the eigenvalues of the autocorrelation and partial autocorrelation matrices of the original data matrix.

A visual method for determining the number of PCs to retain given a specific lag structure is also proposed. The procedure includes assessing the serial correlation in the squared prediction error statistic, Q, together with the cumulative explained variance. Using our method of selecting the number of PCs, we attempt to achieve Q statistics with as low serial correlation as possible and thereby provide a means to monitor the

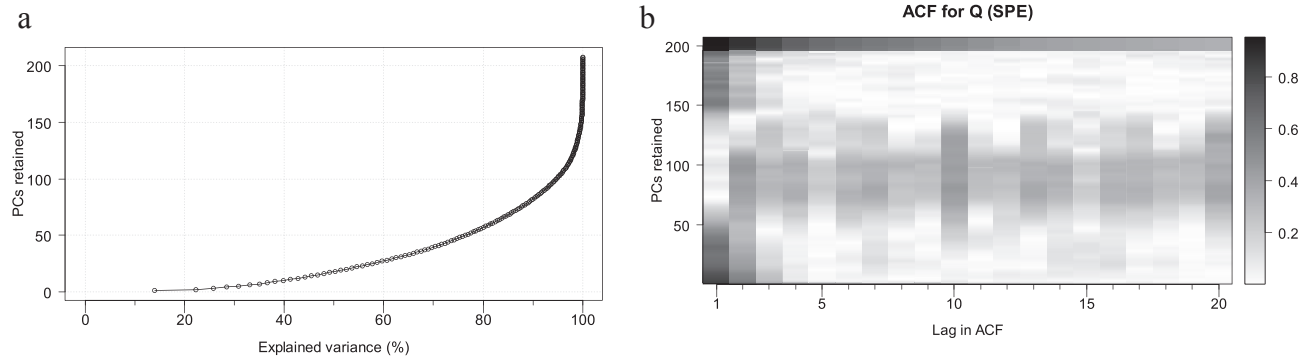

Fig. 12. **a.** Cumulative explained variance given the number of retained PCs in DPCA with 3 lags. **b.** Absolute values of the autocorrelation function for the Q statistics given the number of retained PCs in DPCA with 3 lags.

process dynamics as claimed by Ku et al. [5].

We have illustrated the proposed methods using simulated VAR(1) and VMA(1) processes, and Tennessee Eastman Process data. Our findings indicate that our method is effective in suggesting the proper lag structure and the number of PCs to retain for VAR(1) data. However, the results show that some autocorrelation remains in the Q statistics. Our results also indicate that many lags may be required to model VMA(1) data properly and that it may not be possible to select a number of retained PCs to achieve a low level of autocorrelation in the Q statistics. In fact, for VMA(1) data, our results indicate that static PCA seems to perform well with low levels of autocorrelation in both the Hotelling T^2 and Q statistics. We have also illustrated that in DPCA the inherent issue of adding too many (more than necessary) lags to handle the autocorrelation actually increases the autocorrelation in the monitored statistics. This issue was recently pointed out also in a study by De Ketelaere et al. [7].

Compared to previously proposed methods our method has a more pronounced foundation in multivariate time series theory. An advantage of our method is that it does not assume that any lag needs to be added before considering the amount of autocorrelation in the data. We acknowledge that DPCA normally would be applied to autocorrelated data but we consider it advantageous that our method will indicate that DPCA is not needed when data are in fact time independent. Another advantage of our qualitative approach is that it presents information to the analyst who can use existing process knowledge or additional limitations such as limits on the cumulative explained variance in decision making. When applying DPCA the analyst first needs to decide on how many lags to add. Secondly, given a specific lag structure the number of PCs to retain in the model needs to be decided. A clear drawback of this approach is that it is difficult to automate not least for researchers interested in larger comparative studies.

For analysts interested in reducing the subjective component of the method and make it possible to automate we provide the following recommendations. The maximum lag is set as the lag before which all eigenvalues have dropped below the limit based on the highest eigenvalue from a sample of time-independent data. Such a limit is seen as the dashed horizontal line in the plots of AVEs in this article. The analyst then specifies a lower and upper bound of the cumulative explained variance of the PCA model which the analyst finds acceptable. The number of retained PCs is chosen as to minimize the lag 1 autocorrelation coefficient in the autocorrelation function for the Q statistics while at the same time maintaining the cumulative explained variance within the specified limits.

We view the contribution of this article to be a new method with sound foundation in multivariate time series theory for the modeling step in DPCA before any methods to further reduce the autocorrelation in the monitored statistics or adjust the control limits are applied. In this article, comparisons of our proposed method to available approaches for DPCA are provided using simulated data and data from the TE process. Further studies on comparing the impact of these approaches on process monitoring can be conducted but this was beyond the scope of this article and will be handled as a future research topic.

Acknowledgements

We gratefully acknowledge the financial support from the Swedish Research Council under grant number: 340-2013-5108. We thank two anonymous reviewers and the editor for valuable comments that improved this article.

References

[1] I.T. Jolliffe, Principal Component Analysis, second ed., Springer-Verlag, New York, NY, 2002.
[2] E. Vanhatalo, M. Kulahci, Impact of autocorrelation on principal components and their use in statistical process control, Qual. Reliab. Eng. Int. 32 (4) (2015) 1483–1500, http://dx.doi.org/10.1002/qre.1858.
[3] D.C. Montgomery, C.L. Jennings, M. Kulahci, Introduction to Time Series Analysis and Forecasting, second ed., Wiley, Hoboken, NJ, 2015.
[4] S. Bisgaard, M. Kulahci, Time Series Analysis and Forecasting by Example, Wiley, Hoboken, NJ, 2011.
[5] W. Ku, R.H. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, Chemom. Intell. Lab. Syst. 30 (1995) 179–196.
[6] T. Kourti, J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemom. Intell. Lab. Syst. 28 (1995) 3–21, http://dx.doi.org/10.1016/0169-7439(95)80036-9.
[7] B. De Ketelaere, M. Hubert, E. Schmitt, Overview of PCA-based statistical process monitoring methods for time-dependent, high-dimensional data, J. Qual. Technol. 47 (4) (2015) 1–17.
[8] T.J. Rato, M.S. Reis, Advantage of using decorrelated residuals in dynamic principal component analysis for monitoring large-scale systems, Ind. Eng. Chem. Res. 52 (2013) 13685–13698, http://dx.doi.org/10.1021/ie3035306.
[9] T.J. Rato, M.S. Reis, Defining the structure of DPCA models and its impact on process monitoring and prediction activities, Chemom. Intell. Lab. Syst. 125 (2013) 74–86, http://dx.doi.org/10.1016/j.chemolab.2013.03.009.
[10] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, fifth ed., Prentice Hall, Upper Saddle River, NJ, 2002.
[11] J.E. Jackson, A User's Guide to Principal Components, Wiley, Hoboken, NJ, 2003.
[12] N.D. Tracy, J.C. Young, R.L. Mason, Multivariate control charts for individual observations, J. Qual. Technol. 24 (2) (1992) 88–95.
[13] J.E. Jackson, G.S. Mudholkar, Control procedures for residuals associated with principal component analysis, Technometrics 21 (3) (1979) 341–349.
[14] A. Wachs, D.R. Lewin, Improved PCA methods for process disturbance and failure identification, AIChE J. 45 (8) (1999) 1688–1700.
[15] W. Li, S.J. Qin, Consistent dynamic PCA based on errors-in-variables subspace identification, J. Process Control 11 (2001) 661–678.
[16] G.C. Reinsel, Elements of Multivariate Time Series Analysis, second ed., Springer-Verlag, New York, NY, 1993.
[17] D.C. Montgomery, C.L. Jennings, M. Kulahci, Introduction to Time Series Analysis and Forecasting, Wiley, Hoboken, NJ, 2008.
[18] J.J. Downs, E.F. Vogel, A plant-wide industrial process control problem, Comput. Chem. Eng. 17 (3) (1993) 245–255.
[19] Braatz RD. Massachusetts Institute of Technology, Department of Chemical Engineering. URL: http://web.mit.edu/braatzgroup/links.html. Accessed: 2017-01-16.
[20] P.R. Lyman, C. Georgakis, Plant-wide control of the Tennessee Eastman problem, Comput. Chem. Eng. 19 (3) (1995) 321–331.
[21] E.L. Russel, L.H. Chiang, R.D. Braatz, Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, Chemom. Intell. Lab. Syst. 51 (2000) 81–93.
[22] U. Kruger, Y. Zhou, G.W. Irwin, Improved principal component monitoring of large-scale processes, J. Process Control 14 (2004) 879–888, http://dx.doi.org/10.1016/j.jprocont.2004.02.002.