Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computers and Chemical Engineering 29 (2005) 2355–2362

Computers & Chemical Engineering

www.elsevier.com/locate/compchemeng

Introduction of a nonlinearity measure for principal component models

Uwe Kruger[a,*], David Antory[b], Juergen Hahn[c], George W. Irwin[a], Geoff McCullough[d]

[a] Intelligent Systems and Control Research Group, Queen's University Belfast, BT5 5AH, UK
[b] Virtual Engineering Centre, Cloreen Park, Malone Road, Belfast, BT9 5HN, UK
[c] Department of Chemical Engineering, Texas A&M University, 3122 College Station, TX 77843, USA
[d] Internal Combustion Engines Research Group, Queen's University Belfast, BT9 5AH, UK

Received 5 May 2005; accepted 23 May 2005
Available online 12 July 2005

Abstract

Although principal component analysis (PCA) is an important tool in standard multivariate data analysis, little interest has been devoted to assessing whether the underlying relationship within a given variable set can be described by a linear PCA model or whether nonlinear PCA must be utilized. This paper addresses this deficiency by introducing a nonlinearity measure for principal component models. The measure is based on the following two principles: (i) the range of recorded process operation is divided into smaller regions; and (ii) accuracy bounds are determined for the sum of the discarded eigenvalues. If this sum is within the accuracy bounds for each region, the process is assumed to be linear and vice versa. This procedure is automated through the use of cross-validation. Finally, the paper shows the utility of the new nonlinearity measure using two simulation studies and with data from an industrial melter process.

Keywords: Nonlinearity measure; Principal component analysis; Disjunct regions; Accuracy bounds; Eigenvalues

1. Introduction

For applications in the chemical industry, the reduction aspect of linear PCA is often considered for process modelling (MacGregor, Marlin, Kresta, & Skagerberg, 1991; Wise & Ricker, 1992), monitoring (Kourti & MacGregor, 1996; Raich & Cinar, 1996) and control (Roffel, MacGregor, & Hoffman, 1989; Piovoso & Kosanovich, 1994) of complex processes. Despite the widespread application of linear PCA, complex processes often exhibit nonlinear relationships between the recorded process variables (Jia, Martin, & Morris, 1998; Shao, Jia, Martin, & Morris, 1999) and it is therefore important to determine whether a linear PCA model is sufficiently accurate over the given operating range or whether its nonlinear counterpart has to be applied. However, whenever a nonlinear PCA (NLPCA) model is applied in the literature the process is often assumed to be 'a priori'. To the best of our knowledge, no nonlinearity measure has yet been proposed to determine when to apply, the computationally and conceptually simpler, linear PCA instead of NLPCA. This work addresses this deficiency by introducing a nonlinearity measure for PCA models.

This new measure is based on the analysis of recorded reference data, which cover a predefined range of the process operation and entails the following two principles. The first relates to decomposing this range of operation into small disjunct regions. The second principle relates to the computation of the correlation matrix for each of these disjunct regions, and noting that the elements of this matrix are always obtained using a finite data set. The correlation matrix is fundamental to linear PCA modelling (Jackson, 1991), in terms of defining (i) a reduced dimensional space that represents

* Corresponding author. Tel.: +44 2890 974059; fax: +44 2890 667023.
E-mail address: uwe.kruger@ee.qub.ac.uk (U. Kruger).

0098-1354/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2005.05.013

the score variables, (ii) the complementary space that represents the PCA model residuals, (iii) the variance of the score variables, and (iv) the variance of the PCA model residuals.

Under the assumption that the recorded process variables are stochastic, the estimation of the mean value and variance for each one are characterized by a t- and χ^2-distribution, respectively. This, in turn, implies that 95 or 99% confidence regions can be established for both parameters. The paper demonstrates that these confidence regions can be utilized to determine thresholds for each element in the correlation matrix. Using these thresholds, the paper then shows that maximum and minimum eigenvalues relating to the discarded score variables can be calculated. These in turn allow the determination of both a minimum and a maximum accuracy bound for the variance of the prediction error of the PCA model, since this variance is equal to the sum of the discarded eigenvalues. If this sum lies inside these accuracy bounds for each disjunct region, a linear PCA model is then appropriate over the entire region. Alternatively, if at least one of these sums is outside the accuracy bounds, the error variance of the PCA model residuals then differs significantly for this disjunct region and hence, a nonlinear model is required. The new nonlinearity measure therefore relies on determining whether error variance of the PCA model prediction in each disjunct region is significantly smaller or larger than the uncertainty with which the correlation matrix can be established from the measured data. In order to exclude possible bias from the disjunct region used to obtain the accuracy bounds, a cross-validation principle (Stoica, Eykhoff, Janssen, & Soderstrom, 1986) is used. More precisely, accuracy bounds are obtained for each of the disjunct regions and the sum of the discarded eigenvalues for each is then benchmarked against its corresponding accuracy bounds.

The utility of this new nonlinearity measure is demonstrated using two simple synthetic examples and also data that relate to an industrial chemical process. The two examples involve a linear and a nonlinear relationship between two variables, respectively. For the linear example, the new nonlinearity measure correctly identified that a linear PCA model with one principal component could accurately describe the analyzed range. In contrast, the nonlinearity measure highlighted the fact that the discarded eigenvalue for each region was not within the accuracy bounds for the nonlinear example. Consequently, a NLPCA model was required. Finally, application of the nonlinearity measure to the industrial data set showed that this process is nonlinear although it was previously analyzed using linear PCA (Chen, Wayne, Goulding, & Sandoz, 2000).

2. Preliminaries of principal component analysis

The application of PCA involves the construction of a reduced set of score variables that represent linear combinations of a set of N analyzed variables. The values of the score variables for the ith sample are given by:

$$\mathbf{t}_i = \mathbf{P}^T \mathbf{z}_i, \tag{1}$$

where $\mathbf{t}_i \in \mathbb{R}^n$ is the vector of n score variables or principal components (PCs), $\mathbf{z}_i \in \mathbb{R}^N$ is the vector of the N analyzed variables and $\mathbf{P} \in \mathbb{R}^{N \times n}$ is a transformation matrix containing the first $n < N$ dominant eigenvectors of the correlation matrix $\mathbf{S}_{ZZ} = (1/K-1)\mathbf{Z}^T\mathbf{Z} \in \mathbb{R}^{N \times N}$ as column vectors. $\mathbf{S}_{ZZ}$ is calculated using a reference data set $\mathbf{Z}^T = [\,\mathbf{z}_1 \quad \mathbf{z}_2 \quad \cdots \quad \mathbf{z}_i \quad \cdots \quad \mathbf{z}_K\,]$ of K mean-centered observations, which are scaled to unit variance. The mismatch error between the measured and predicted sensor readings, $\mathbf{e}_i$, is:

$$\mathbf{e}_i = \mathbf{z}_i - \mathbf{P}\mathbf{t}_i = [\mathbf{I} - \mathbf{P}\mathbf{P}^T]\,\mathbf{z}_i, \tag{2}$$

The application of PCA projects $\mathbf{z}_i$ onto a model plane and a residual subspace. The model plane is spanned by the first n dominant eigenvectors of $\mathbf{S}_{ZZ}$ and describes the linear relationships between the analyzed variables. In contrast, the residual subspace is spanned by the remaining $(N-n)$ eigenvectors of $\mathbf{S}_{ZZ}$ and represents the mismatch error of the PCA prediction.

3. Nonlinearity measure for principal component models

This section introduces a new nonlinearity measure for use with principal component models. The measure relies on the division of the recorded range of process operation, or operational range, into disjunct regions. A correlation matrix is then obtained using the data of one of these regions. This is followed by computing thresholds for each matrix element on the basis of the confidence limits for computing the mean and variance of each process variable. Then, the maximum and minimum sum of discarded eigenvalues, or the accuracy bounds, are calculated using the fact that the matrix elements are within known thresholds. Finally, the sum of discarded eigenvalues are obtained for the correlation matrix of each region, noting that the mean and variance of each process variable are obtained for the region for which the accuracy bounds are obtained.

If the sum of discarded eigenvalues for the PCA model of each region is inside the accuracy bounds, the process is said to be linear. Conversely, if at least one of these sums is outside, the process must be assumed to be nonlinear. In order to remove a possible bias from the region that is used to obtain the accuracy bounds, a cross-validation principle is utilized here. This implies that accuracy bounds are obtained for each region and the sum of discarded eigenvalues are benchmarked against their corresponding bounds. The remainder of this section is organized as follows. The assumptions for applying this new measure are given in Section 3.1. Section 3.2 then discusses issues relating to the division of the operational range. After division, a PCA model is then identified for each disjunct region. Section 3.3 shows how to determine thresholds for each element of the correlation matrix and Section 3.4 outlines how to establish accuracy bounds for the sum of the discarded eigenvalues. Section 3.5 then defines the procedure for calculating the new nonlinearity measure, which relies on evaluating whether the differences in the error variance of the PCA model prediction in each disjunct region is significantly smaller, or larger, than the uncertainty with which the correlation matrix can be obtained. Finally, Section 3.6 contains two simple examples.

3.1. Assumptions

The new nonlinearity measure is based on the following assumptions.

Assumption 1. The variables are mean-centered and scaled to unit variance with respect to disjunct regions for which the accuracy bounds are to be determined.

Assumption 2. Each of these disjunct regions contains the same number of process observations.

Assumption 3. A PCA model is available for one region, where the accuracy bounds for the sum of the discarded eigenvalues can then be obtained.

Assumption 4. PCA models are obtained for the remaining disjunct regions.

Assumption 5. The same number of principal components is retained in each of the PCA models.

3.2. Defining disjunct regions

The division of the original operating range into disjunct regions can either be accomplished by utilizing a priori knowledge of the process, for example, or by direct analysis of the recorded data. The former approach could use knowledge about distinct operating regions of the plant. The latter approach could entail a PCA analysis to identify distinct operating regions as discussed in Wold, Esbensen, and Geladi (1987). If this, however, does not reveal any distinctive features, the original operating region could initially be divided into two disjunct regions, with the nonlinearity measure applied as discussed below. The impact of increasing the number of disjunct regions is then examined. Note, however, that increasing the number of disjunct regions reduces the number of samples in each, an issue elaborated upon later in the next section. One potential remedy is to collect a large enough set of reference data so that the size of the data set in each disjunct region is sufficient to avoid large uncertainties arising in the elements of the correlation matrix.

A remaining question is which of the disjunct regions should be used to establish the accuracy bounds? One might consider using the most centered region or alternatively, one at the margins of the original operational range. Practically, the region at which the process is most often expected to operate could be chosen. The principle of cross-validation (Stone, 1974; Stoica et al., 1986) can usefully be employed to automate this selection. Wold (1978) and Krzanowski (1983) advocated the use of cross-validation for determining the number of retained principal components. The use of cross-validation entails the calculation of accuracy bounds for each disjunct region. If such accuracy bounds are initially determined for a region at one of the margins, for example, the sum of discarded eigenvalues for each successive region is benchmarked against these. A second set of accuracy bounds is then determined for the region next to the first one and the sum of discarded eigenvalues of each successive region is then benchmarked against these and so on. This procedure is carried out until accuracy bounds for each of the disjunct regions have been obtained along with the sum of discarded eigenvalues for all the rest. Note that the PCA models will vary from region to region. This results from the normalization procedure, since the mean and variance of each variable may change depending on which region is currently being analyzed.

3.3. Thresholds for the elements of the correlation matrix

The original operating range is divided into a total of m disjunct regions, each of which includes a total of $\tilde{K} = K/m$ observations rounded to the nearest integer. The correlation matrix of the N analyzed variables for one of the disjunct regions has the following structure:

$$\mathbf{S}_{ZZ} = \begin{bmatrix} s_{11} & s_{11} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{bmatrix}. \tag{3}$$

This matrix is symmetric and the elements are defined as:

$$s_{ij}\frac{1}{\tilde{K}-1}\sum_{k=1}^{\tilde{K}}\frac{z_{ki}-\bar{z}_i}{\sigma_i}\frac{z_{kj}-\bar{z}_j}{\sigma_j}, \tag{4}$$

where $\bar{z}_i, \bar{z}_j$ are the estimated mean values and σ_i, σ_j are the estimated standard deviations of the ith, jth variables, respectively.

Since $\tilde{K}$ is finite, the estimates of the mean and variance follow t- and χ^2-distributions, respectively. The confidence limits of the ith mean value, ${}_z\mathrm{CONF}_\alpha^{(i)}$, for a confidence level of $\alpha = 95\%$ or $\alpha = 99\%$ can be determined as shown in Table 1. The procedure for determining the confidence regions for the variance of the ith variable, ${}_s\mathrm{CONF}_\alpha^{(i)}$ is summarized in Table 2. Incorporating the confidence intervals of the mean

Table 1
Calculation of confidence limits for the mean value of the *i*th variable

Step	Description	Equation
1	Determine solution of c_i for a *t*-distribution $f_1\,(\cdot)$	$c_i = f_1^{-1}\left(\frac{1+\alpha}{2}\right)$
2	Calculate mean, $\tilde{z}_i$, and variance, s_i, of $z_{1i},\ \ldots,\ z_{\tilde{K}i}$	$\tilde{z}_i = \frac{1}{\tilde{K}}\sum_{k=1}^{\tilde{K}} z_{ki}, \qquad s_i = \frac{1}{\tilde{K}-1}\sum_{k=1}^{\tilde{K}}(z_{ki}-\tilde{z}_i)^2$
3	Compute μ_i	$\mu_i = \frac{s_i c_i}{\sqrt{\tilde{K}}}$
4	Define confidence limit as follows	${}_z\text{CONF}_\alpha^{(i)}\{\tilde{z}_i - \mu_i \le \hat{\bar{z}}_i \le \bar{z}_i + \mu_i\}$

Table 2
Calculation of confidence limits for the variance of the *i*th variable

Step	Description	Equation
1	Determine solution of c_{1i} and c_{2i} for χ^2-distribution $f_2\,(\cdot)$	$c_{1i} = f_2^{-1}\left(\frac{1-\alpha}{2}\right), \qquad c_{2i} = f_2^{-1}\left(\frac{1+\alpha}{2}\right)$
2	Calculate $(\tilde{K}-1)\,s_i$	
3	Compute μ_{1i} and μ_{2i}	$\mu_{1i} = \frac{(\tilde{K}-1)\,s_i}{c_{1i}}, \qquad \mu_{2i} = \frac{(\tilde{K}-1)\,s_i}{c_{2i}}$
4	Define confidence limit as follows	${}_s\text{CONF}_\alpha^{(i)}\{\mu_{2i} \le \hat{s}_i \le \mu_{1i}\}$

and variance, an upper and lower threshold can be determined for each element of the correlation matrix. More precisely, Eq. (3) can be rewritten as follows:

$$\mathbf{S}_{ZZ} = \begin{bmatrix} s_{11_U} \le s_{11} \le s_{11_L} & s_{12_U} \le s_{12} \le s_{12_L} & \cdots & s_{1N_U} \le s_{1N} \le s_{1N_L} \\ s_{21_U} \le s_{21} \le s_{21_L} & s_{22_U} \le s_{22} \le s_{22_L} & \cdots & s_{2N_U} \le s_{2N} \le s_{2N_L} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1_U} \le s_{N1} \le s_{N1_L} & s_{N2_U} \le s_{N2} \le s_{N2_L} & \cdots & s_{NN_U} \le s_{NN} \le s_{NN_L} \end{bmatrix}, \tag{5}$$

where the indices *U* and *L* refer to the upper and lower threshold, respectively. A simplified version of Eq. (5) is given below.

$$\mathbf{S}_{ZZ_U} \le \mathbf{S}_{ZZ} \le \mathbf{S}_{ZZ_L}, \tag{6}$$

where the matrices $\mathbf{S}_{ZZ_U}$ and $\mathbf{S}_{ZZ_L}$ contain the values that represent the upper and lower thresholds, respectively.

The confidence limits for the mean and variance of each variable depend on the number of samples; the smaller the sample set the wider the confidence limits. It is therefore necessary to have a sufficiently large reference set from the analyzed process (i) to guarantee that the sample sets of adjacent regions after subdivision does not produce wide confidence limits, (ii) enable the generation of enough disjunct regions and (iii) provide enough information about the process.

3.4. Determination of accuracy bounds

This subsection relies on the fact that the sum of the residual variance, σ, is equal to the sum of the discarded eigenvalues of a PCA model:

$$\sigma = \sum_{j=1}^{N}\sigma_j = \frac{1}{\tilde{K}-1}\sum_{i=1}^{\tilde{K}}\sum_{j=1}^{N} e_{ij}^2 = \sum_{k=n+1}^{N}\lambda_k, \tag{7}$$

where σ_j represents the residual variance for the prediction of the *j*th variable, $e_{ij} = z_{ij} - \mathbf{z}_i^T\mathbf{P}\mathbf{p}_j$ with $\mathbf{p}_j$ being the *j*th row vector of **P** and λ_k the *k*th largest eigenvalue of $\mathbf{S}_{ZZ}$.

Since the eigenvalues $\lambda_{n+1}, \ldots, \lambda_N$ depend on the elements of the correlation matrix $\mathbf{S}_{ZZ}$, a set of eigenvalues can be obtained for which each eigenvalue is as large as possible. Correspondingly a set can be obtained for which each eigenvalue is as small as possible. This gives rise to the following optimization problem:

$$\begin{aligned} \lambda_{k_{\text{MAX}}} &= arg \max_{\Delta\mathbf{S}_{ZZ_{\text{MAX}}}} \lambda_k(\mathbf{S}_{ZZ} + \Delta\mathbf{S}_{ZZ_{\text{MAX}}}) \\ \lambda_{k_{\text{MIN}}} &= arg \min_{\Delta\mathbf{S}_{ZZ_{\text{MIN}}}} \lambda_k(\mathbf{S}_{ZZ} + \Delta\mathbf{S}_{ZZ_{\text{MIN}}}) \end{aligned}, \tag{8}$$

which is subject to the following constraints:

$$\begin{aligned} \mathbf{S}_{ZZ_U} \le \mathbf{S}_{ZZ} + \Delta\mathbf{S}_{ZZ_{\text{MAX}}} \le \mathbf{S}_{ZZ_L} \\ \mathbf{S}_{ZZ_U} \le \mathbf{S}_{ZZ} + \Delta\mathbf{S}_{ZZ_{\text{MIN}}} \le \mathbf{S}_{ZZ_L} \end{aligned}. \tag{9}$$

Here $\Delta\mathbf{S}_{ZZ_{\text{MAX}}}$ and $\Delta\mathbf{S}_{ZZ_{\text{MIN}}}$ are perturbations of $\mathbf{S}_{ZZ}$ that produce a maximum value, $\lambda_{k_{\text{MAX}}}$ and minimum value, $\lambda_{k_{\text{MIN}}}$ of λ_k, respectively. This maximum value for each discarded eigenvalue allows, according to Eq. (7), the estimation of a maximum threshold, σ_{MAX}, and a minimum threshold, σ_{MIN}, for the sum of the residual variances of the PCA model prediction:

$$\sigma_{\text{MAX}} = \sum_{k=n+1}^{N}\lambda_{k_{\text{MAX}}}, \qquad \sigma_{\text{MIN}} = \sum_{k=n+1}^{N}\lambda_{k_{\text{MIN}}}. \tag{10}$$

These maximum and minimum thresholds are further defined as accuracy bounds. This relates to the fact that any set of reference data from the same process in the same operating region cannot produce a larger or a smaller variance of the PCA model prediction.

A genetic optimization strategy that relies on (Sharma & Irwin, 2003) was considered for determining the maximum and minimum solution for λ_k. This incorporated a fuzzy coding method and offered a faster convergence for real-valued parameters compared to alternative conventional encoding methods. The genetic strategy was set to run for 2000 generations at each stage and the best chromosome represented the corresponding optimum set of parameters, which included the optimum values of $\lambda_{k_{\mathrm{MAX}}}$ and $\lambda_{k_{\mathrm{MIN}}}$ and the unknown parameters in $\Delta\mathbf{S}_{ZZ_{\mathrm{MAX}}}$ and $\Delta\mathbf{S}_{ZZ_{\mathrm{MIN}}}$. The algorithm here included 20 chromosomes with a crossover probability $p_c = 0.65$ and a mutation probability $p_m = 0.01$.

3.5. Definition of the new nonlinearity measure

The accuracy bounds defined above are determined for only one of the disjunct regions and therefore define the maximum and minimum accuracy for predicting the analyzed variable set for that region. If a PCA model is produced for all of the disjunct regions, a set of discarded eigenvalues, λ_k, is then available for each. The sum of the discarded eigenvalues for each region set is then benchmarked against the available accuracy bounds. If all of these sums fall inside the accuracy bounds, the process is said to be linear. Alternatively, if one lies outside the accuracy bounds, it must be concluded that the residual variance for this region, defined in Eq. (7), is smaller or larger than the effect of the uncertainty in the PCA model accuracy.

Hence, a linear PCA model would give a different prediction accuracy for this region, implying that a nonlinear one must be employed.

However, this approach relies only on the analysis of single adjacent regions and the results may vary depending on which of the adjacent regions is used. The question then arises as to which of the adjacent regions should be selected. Practically, the region where the process is most often expected to operate may be chosen. To remove possible bias from the region used to obtain the accuracy bounds, the principle of cross-validation (Stone, 1974; Stoica et al., 1986) is utilized in this work. More precisely, accuracy bounds are obtained for each of the adjacent regions, noting that the mean and variance of each variable may change with the region selected. Consequently, the PCA model for each region may also change.

The sum of the discarded eigenvalues for each of the disjunct regions is then benchmarked against their corresponding accuracy bounds. The process is said to be linear if, and only if, this sum lies inside its own regional accuracy bounds and those for the remaining disjunct regions. Conversely, if at least one sum of discarded eigenvalues falls outside an accuracy bound, the process must be considered to be nonlinear. This follows because the uncertainty with which a linear PCA model can predict a disjunct region of the process is smaller than, or larger than, the variance of the model prediction error for the region for which the accuracy bounds have been computed. Two simple application studies to demonstrate the utility of the new nonlinearity measure are described next.

3.6. Simple examples

The two examples are based on an analysis of two variables, one with a linear relationship between the variables, the other a nonlinear relationship. The aim is to demonstrate the utility of the new nonlinearity measure, and so accuracy bounds for only one of the disjunct regions were obtained.

3.6.1. Linear example

The linear example is based on a random variable x for which a total of 3000 normally distributed samples, scaled between −3 and 3 were obtained. The two analyzed variables, z_1 and $z_2$were then generated as follows:

$$z_1 = x + 0.05e_1, \qquad z_2 = x + 0.05e_2, \tag{11}$$

where e_1 and e_2 are independently and identically distributed sequences of zero mean and unit variance, representing normally distributed measurement uncertainty. A scatter diagram for z_1 and z_2 is shown in Fig. 1. The set of reference data was then subdivided into three disjunct regions, each contained 1000 samples. This division is shown graphically in Fig. 2, where only 100 samples for each of the disjunct regions are depicted to enhance the presentation. After mean-centering and scaling the data set 1 to unit variance, accuracy bounds were found for set 1 as described earlier. Both variables, z_1 and z_2, were highly correlated and hence, only the first principal component needed to be retained for PCA modelling. The second principal component in this case described the influence of measurement uncertainty, and hence, the variance of the prediction error of the PCA model. The limits on the accuracy bound, computed for confidences of 95 and 99%,

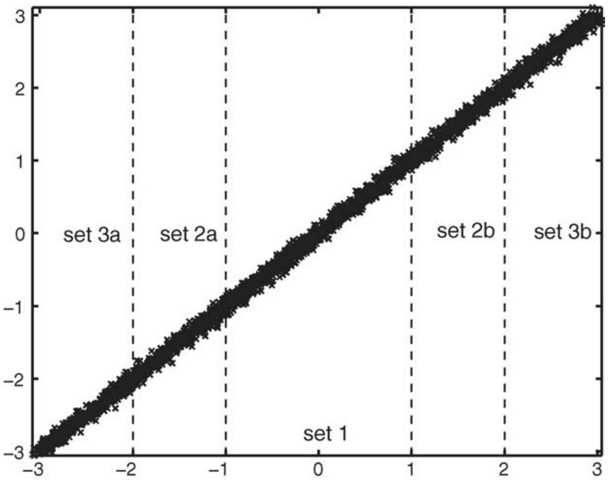

Fig. 1. Original scatter diagram of simulated data (linear example).

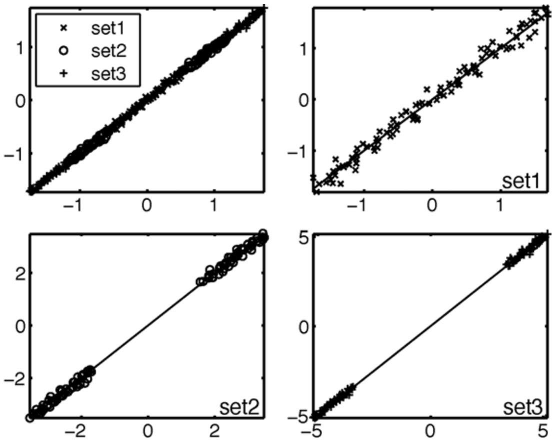

Fig. 2. Division of original operating range into three disjunct regions.

are given in Table 3. This table also includes the value of the discarded eigenvalue of the correlation matrices for data set 2 and data set 3. The discarded eigenvalues of the correlation matrix for all three data sets fall inside the accuracy bounds since the process is linear. This is graphically illustrated by the upper plot in Fig. 3.

3.6.2. *Nonlinear example*

The nonlinear example was also based on the generation of 3000 samples of a normally distributed variable x in the range between −3 and 3 with zero mean. The two analyzed variables were then computed as follows:

$$z_1 = x + 0.05e_1, \qquad z_2 = \sin x + 0.05e_2. \tag{12}$$

Fig. 4 shows the scatter diagram for the variables z_1 and z_2. This set of reference data was then divided into 3 disjunct regions containing 1000 samples each, as illustrated in Figs. 4 and 5. In this example, the accuracy bounds were calculated for the portion of the original operational range falling in the “middle” of the scatter plot, i.e., data set 1. Note that the two other disjunct regions, data set 2 and data set 3, are divided into two segments, where data set 2 consists of data set 2a and data set 2b and data set 3 consists of data set 3a and data set 3b, respectively. Table 4 gives the accuracy bounds, for confidence limits of 95 and 99%. According to Eq. (12), both variables depend on x and so, only the first principal component was retained for the PCA model of each disjunct region. Again, a second principal component described the PCA model prediction error for each region.

Table 3
Nonlinearity measure applied to simulation data of the linear process

UCL99	0.0083
UCL95	0.0081
LCL95	0.0068
LCL99	0.0066
Region$_1$	0.0074
Region$_2$	0.0071
Region$_3$	0.0072

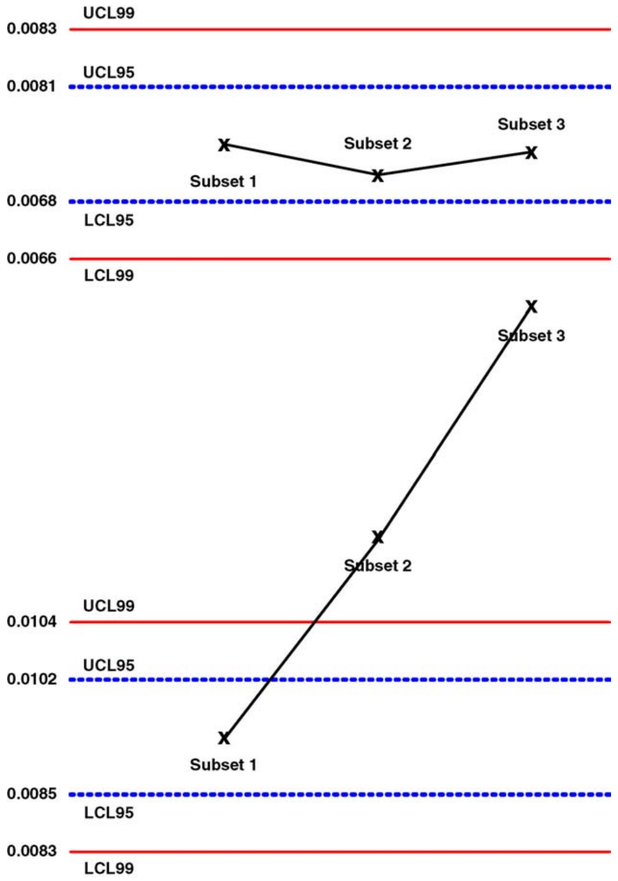

Fig. 3. Graphical representation of nonlinearity measure for three disjunct regions.

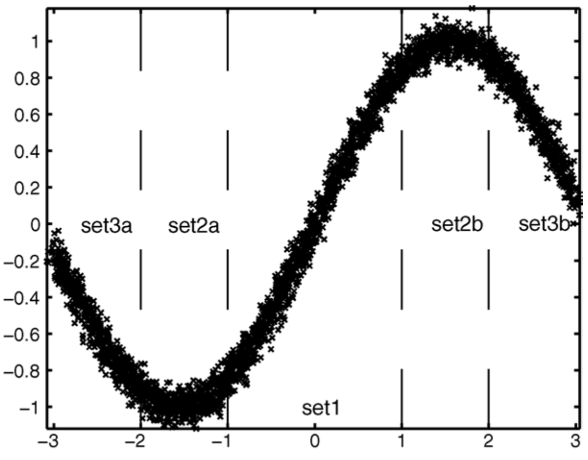

Fig. 4. Original scatter diagram of simulated data (nonlinear example).

Table 4
Nonlinearity measure applied to simulation data of nonlinear process

UCL99	0.0104
UCL95	0.0102
LCL95	0.0085
LCL99	0.0083
Region$_1$	0.0093
Region$_2$	0.0767
Region$_3$	0.2990

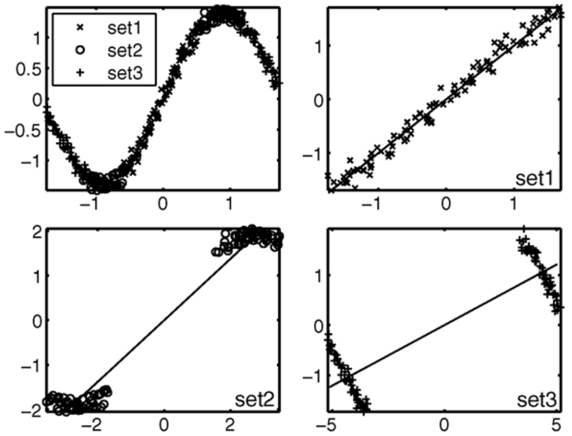

Fig. 5. Division of original operating range into three disjunct regions.

The second discarded eigenvalue for each of region is also given in Table 5. In contrast to the linear example, the lower plot in Fig. 2 shows that the discarded eigenvalue of the PCA model for data set 2 and data set 3 fell outside the accuracy bounds. This highlights the fact that a linear PCA model is inappropriate for this example. The new nonlinearity measure has thus clearly indicated the need for a NLPCA model.

4. Industrial data set

This section presents the application of the new nonlinearity measure to data from an industrial melter process. The process is briefly described first, followed by a report of the degree of nonlinearity within the recorded data set.

4.1. Process description (Chen et al., 2000)

The melter process is part of a disposal procedure. Waste material is preprocessed by an evaporation treatment to produce a powder, which is then clad by glass provided by the melter process. The melter vessel is continuously filled with powder and raw glass is discretely introduced in the form of glass frit. This binary composition is heated by four induction coils, which are positioned around the vessel. Because of the heating procedure, the glass is homogeneously melted. The process of filling and heating continues until the desired height of the liquid column is reached. Then, the molten mixture is poured out through an exit funnel. After the contents of the vessel are emptied to the height of the nozzle, the next cycle of filling and heating is carried out. Measurements of eight temperatures, the power in four induction coils, the viscosity of the molten glass and voltage were taken every 5 min. The process therefore presented 14 variables for inclusion in the PCA analysis.

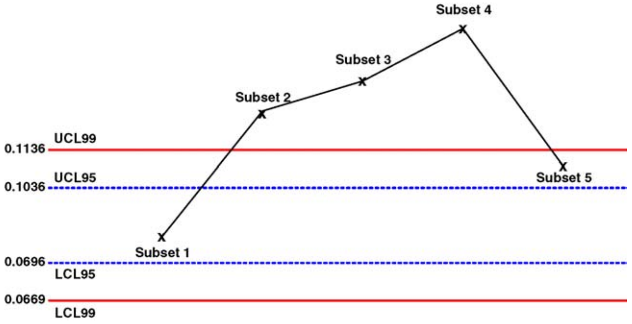

Fig. 6. Graphical representation of nonlinearity measure for accuracy bounds for first of five disjunct regions.

4.2. Degree of process nonlinearity

A sample set including 1000 samples was recorded from the melter process. This set was divided into five disjunct regions of 200 samples each. Note that the division into substantially more than five regions would have led to larger thresholds for the elements of the correlation matrix. The results of applying the new nonlinearity measure for determining the accuracy bounds for each region are summarized in Table 5. This shows the upper and lower boundaries of the accuracy bounds for each disjunct region, and also the sum of the discarded eigenvalues. These limits were based on the thresholds for each element of the correlation matrix corresponding to confidence levels of 95 and 99% in determining the mean and variance for each process variable. Note that the process variables were normalized with respect to the mean and variance of the regions for which the accuracy bounds were computed.

The sum of the discarded eigenvalues for the region in which the accuracy bounds were obtained, indicated by italics in Table 5, were, as expected, inside these bounds. In contrast, at least one of these sums, indicated by the bold numbers in Table 5, lay outside any of the accuracy bounds. Fig. 6 is a graphical representation of the case where the accuracy bounds were obtained for the first disjunct region. Whilst the sum of the discarded eigenvalues for the first and fifth disjunct region were inside the accuracy bounds, those

Table 5
Nonlinearity measure applied to melter process using five subsets

Region	UCL99	UCL95	LCL95	LCL99	Region$_1$	Region$_2$	Region$_3$	Region$_4$	Region$_5$
1	0.1136	0.1036	0.0696	0.0669	*0.0798*	**0.1253**	**0.1432**	**0.1719**	0.1072
2	0.1263	0.1153	0.0774	0.0746	**0.0585**	*0.0904*	0.1062	**0.1259**	0.0799
3	0.1348	0.1261	0.0847	0.0794	0.1034	**0.0574**	*0.0898*	0.1248	0.0794
4	0.2722	0.2561	0.1721	0.1601	0.212	**0.1019**	0.1674	*0.1834*	**0.1333**
5	0.1447	0.1249	0.0852	0.0839	0.0889	**0.0633**	0.0997	0.1139	*0.1372*

for the remaining regions fell outside these bounds, indicating a larger residual variance of the PCA model prediction for those regions. This, in turn, implied that the melter process is nonlinear, as the variance of the PCA model residuals across the original operating region was larger than could be explained by the uncertainty in determining the correlation matrix. Consequently, a NLPCA model is required for this process. It should also be noticed that the selection of the region for which the accuracy bounds are calculated is important. More precisely, the number of sums of discarded eigenvalues that violate the accuracy bounds varied from only one to three. However, the automated procedure, involving cross-validation, was capable of overcoming this problem, as all of the disjunct regions are considered in determining the accuracy bounds.

5. Summary and conclusions

This paper introduced a new nonlinearity measure for use with principal component models. This is based on two principles, one relating to the division of the original operational range described by the recorded reference data into a number of disjunct regions. The second principle uses the fact that the correlation matrix, which is fundamental in determining a PCA model, is computed from a finite data set. This lends itself to establish confidence regions for the mean and variance of each process variable. Consequently, thresholds can be computed for each element of the correlation matrix.

These thresholds are then utilized to compute a sum of discarded eigenvalues that represent a maximum and a minimum value, which represent accuracy bounds for the PCA model prediction. If the sum of the PCA model of each region is inside these accuracy bounds, the process is said to be linear. Conversely, if one of these sums falls outside, the process requires a NLPCA model. To exclude possible bias from the region used to obtain the accuracy bounds, the cross-validation principle was used in this work.

The utility of the new nonlinearity measure was demonstrated using two simple examples, the first describing a linear relationship and the second presenting a nonlinear relationship between two variables. Whilst the nonlinearity measure suggested that a linear model is sufficient for the first one, the second example produced violations of the accuracy bounds, hence, correctly suggesting that a NLPCA model is required. The paper finally presented the application of the new nonlinearity measure to an industrial example. The division of the original operational range into five disjunct regions revealed that the process variables described a nonlinear relationship between them, requiring the use of a NLPCA model.

References

Chen, Q., Wayne, R., Goulding, P. R., & Sandoz, D. J. (2000). The application of principal component analysis and kernel density estimation to enhance process monitoring. *Control Engineering Practice*, *8*(5), 531–543.

Jackson, J. E. (1991). A users guide to principal components. In *Wiley Series in Probability and Mathematical Statistics*. New York: John Wiley.

Jia, F., Martin, E. B., & Morris, A. J. (1998). Non-linear principal component analysis for process fault detection. *Computers and Chemical Engineering*, *22*, S851–S854.

Kourti, T., & MacGregor, J. F. (1996). Multivariate spc methods for process and product management. *Journal of Quality Technology*, *28*, 409–428.

Krzanowski, W. J. (1983). Cross-validatory choice in principal component analysis: Some sampling results. *Journal of Statistical Computation and Simulation*, *18*, 299–314.

MacGregor, J. F., Marlin, T. E., Kresta, J. V., & Skagerberg, B. (1991). Multivariate statistical methods in process analysis and control. In *AIChE Symposium Proceedings of the Fourth International Conference on Chemical Process Control, No. P-67* (pp. 79–99). New York: AIChE Publication.

Piovoso, M. J., & Kosanovich, K. A. (1994). Applications of multivariate statistical methods to process monitoring and controller design. *International Journal of Control*, *59*(3), 743–765.

Raich, A., & Çinar, A. (1996). Statistical process monitoring and disturbance diagnosis in multivariable continuous processes. *AIChE Journal*, *42*(4), 995–1009.

Roffel, J. J., MacGregor, J. F., & Hoffman, T. W. (1989). The design and implementation of a multivariable internal model controller for a continuous polybutadiene polymerisation train. In *Proceedings of the IFAC Conference on Dynamics and Control of Chemical Reactors* (pp. 9–15).

Shao, R., Jia, F., Martin, E. B., & Morris, A. J. (1999). Wavelets and nonlinear principal component analysis for process monitoring. *Control Engineering Practice*, *7*(7), 865–879.

Sharma, S. K., & Irwin, G. W. (2003). Fuzzy coding of genetic algorithms. *IEEE Transactions on Evolutionary Computations*, *7*(4), 344–355.

Stoica, P., Eykhoff, P., Janssen, P., & Soderstrom, T. (1986). Model structure selection by cross-validation. *International Journal of Control*, (6), 1841–1878.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction (with discussion). *Journal of the Royal Statistical Society, Series B Methodological*, *36*, 111–133.

Wise, B. M., & Ricker, N. L. (1992). Identification of finite impulse response models by principal component regression. *Process Control and Quality*, *4*, 77–86.

Wold, S. (1978). Cross-validatory estimation of the number of principal components in factor and principal component models. *Technometrics*, *20*(4), 397–406.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*, 37–52.