Global Advanced Research Journal of Agricultural Science (ISSN: 2315-5094) Vol. 4(10) pp. 711-724, October, 2015.
Available online http://garj.org/garjas/home

Full Length Research Paper

Nonlinear Fuzzy Robust PCA for Fault Detection of Environmental Processes

Majdi Mansouri[a], Marie-France Destain[b], Hazem Nounou[a], Mohamed Nounou[c]

[a] Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha, Qatar,
[b] University of Liege - Gembloux Agro-Bio Tech Faculty Department of Biosystems Engineering, Gembloux, Belgium.
[c] Chemical Engineering Program, Texas A&M University at Qatar, Doha, Qatar.

Accepted 20 October, 2015

Fault detection is often utilized for proper operation of environmental processes. In this paper, a nonlinear statistical fault detection using nonlinear fuzzy robust principal component analysis (NFRPCA) -based generalized likelihood ratio test (GLRT) is proposed. The objective of this work is to extend our previous work (Mansouri et al., 2015), to achieve further improvements and widen the applicability of the developed method in practice by using the NFRPCA method. It is well known that the principal components are often affected by outliers, thus may not capture the true structure of the data. Therefore data reduction based on PCA becomes unreliable if outliers are present in the data. To relieve the noise sensitivity, to obtain accurate principal components of a data, and to reduce the effective system dimension, we propose to use the nonlinear fuzzy robust principal component analysis. The objective of this paper is to combine the GLRT with NFRPCA model in an attempt to improve the performance of fault detection. GLRT-based NFRPCA is a multivariate statistical method utilized in fault detection. Here the fault detection problem is addressed so that the data are first modeled using the NFRPCA analysis algorithm and then the faults are detected using generalized likelihood ratio test. The data is collected from the crop model data in order to calculate the NFRPCA model, the thresholds and the fault detection indices (Hotellingstatistic T^2, Q statistic). It is demonstrated that the performance of faults detection can be improved by combining GLRT and NFRPCA.

Keywords: Environmental processes, fault detection, Generalized likelihood ratio test, Nonlinear fuzzy robust, Principal component analysis.

INTRODUCTION

Effective operation of various engineering systems requires tight monitoring of some of their key process variables. Process systems are using large amount of data from many variables that are monitored and recorded continuously every day. For these reasons, the problem of fault detection that responses effectively to faults that mislead the process and harm the system reliability represents a key process in such operation of these systems(Mourad and Bertrand-Krajewski 2002).Several multivariate statistical techniques for fault detection, analysis of process and diagnosis have been developed and used in practice. These techniques are useful since operation safety

*Corresponding Author's E-mail: majdi.mansouri@qatar.tamu.edu; Tel:+974.7773.4583; Fax: +974.4423.0065.

and the better quality products are some of the main goals in the industry applications. Faults detection has been performed manually using data visualization tools (Tang et al., 2001), however these tools takes a lot of time for real-time detection with continuous data. In the most recent years, researchers have proposed machine learning and automated statistical methods like: nearest neighbor (Ramaswamy et al., 2000; Bolton et al., 2001), clustering (Rousseeuw and Ruts 1996), minimum volume ellipsoid (Ruts and Rousseeuw 1996), convex pealing (Gonzalez et al., 2002), neural network classifier (John 1995), decision tree (Bulut et al., 2005) and support vector machine classifier (Jackson and Mudholkar 1979). These proposed techniques are quicker than other manual techniques, however there are disadvantages which make them inadequate for continuous fault detection for the cases of streaming data. More recently, principal component analysis (PCA) and multivariate statistical process control(MSPC) approach are proposed to overcome these problems. The authors in (Chiang et al., 2001), have proposed PCA as a tool of MSPC. Also, PCA was defined as a method which projects a high dimensional measurement space into a lower dimensional space (Mac Gregor and Kourti 1995). PCA provides linear combinations of parameters which demonstrate most common trends in a data set. In mathematical terms, PCA relies on the orthogonal decomposition of the covariance matrix over the process variables along with the directions which give the maximum data variation. It is also mentioned that PCA is researched for two problems: the MSPC (David and Marta 2008), and fault detection and isolation (FDI) problem (Luukka 2011). Authors of (Luukka 2011), have listed diagnosis and fault detection techniques in three categories: (i) quantitative model-based schemes, (ii)qualitative model schemes and corresponding search strategies and (iii) process data based techniques. PCA falls into the third category since it utilizes databases in an attempt to obtain the statistical (PCA model). The main indices used with PCA methods are Hotellingstatistic, T^2 ; sum of squared residuals, SPE; and/or Q statistics. The T^2 statistic isa way to measure the variation captured in the PCA model whereas the Q statistic is a way to measure the amount of variation which was not captured by the PCA model. PCA is known to be one of the most popular MSPC monitoring methods. Nevertheless, there are some disadvantages of it. One disadvantage is that the PCA is not suitable for monitoring processes that show non-stationary behavior. The other shortcoming of the PCA model is that most of the processes run under different circumstances. The use of standard PCA solution in this kind of processes might produce too many missed faults, since the grade transitions from one operation mode to another operation mode might damage the correlation existing between various parameters. In addition, the disturbances that are measured may be treated as faults. Moreover, the principal components resulted from the standard PCA are often affected by outliers or noise and may not capture the true structure of the data, we propose to use the nonlinear fuzzy robust principal component analysis (NFRPCA) model in order to reduce the noise sensitivity, to obtain accurate principal components of a set of data and to deal with the nonlinearity in variables. The NFRPCA model decreases the dimensionality of the original data by projecting it into a space with significantly fewer dimensions. It results in the principal events of nonlinear variability in a process. If any of the events change, it might be a result of a fault in the process. Moreover, NFRPCA uses a fuzzy covariance matrix instead of the traditional data covariance matrix in order to relieve the noise sensitivity and produces the diagnostic statistics based on the influence function. In the current work, we address problem of the fault detection in environmental processes representing the crop model so that the data are first modeled using the NFRPCA analysis algorithm and then the faults are detected using generalized likelihood ratio test. Generalized likelihood ratio (GLRT)-based NFRPCA is proposed to detect the faults when the data are first modeled with the NFRPCA. This NFRPCA presented here is derived from the nonlinear case of fuzzy principal component analysis algorithm introduced in (Gustafsson 1996) and it is investigated here as modeling algorithm in the task of fault detection (Nguyen and Widrow 1990; Willsky et al., 1980). The NFRPCA is used to create the model and find nonlinear combinations of parameters which describe the major trends in a data set and GLRT is used to detect the faults and both are utilized to improve faults detection process. GLRT has been proposed in order to establish an adaptive system, which reaches three important problems; estimation, fault detection and magnitude compensation of jumps. GLRT is proposed for fault detection of different applications: geophysical signal segmentation (Willsky et al., 1980), signals and dynamic systems (Nguyen and Widrow 1990), incident fault detection on freeways (Dawdle et al., 1982), missiles trajectory (Tamura and Tsujita 2007). Therefore, in the current work it is proposed to benefit from the advantages of the GLRT (Mourad and Bertrand-Krajewski 2002), in order to improve the fault detection task in the cases where process model is not available. It is also compared to classical NFRPCA indices T^2 and Q.

The rest of the paper is organized as the following. In Section 2, an introduction to NFRPCA is given, followed by descriptions of the two main detection indices, T^2 and Q, which are generally used with NFRPCA for fault detection. Then, the GLRT which is utilized in composite hypothesis testing is discussed in Section 3. After that, the NFRPCA based GLRT method used for detecting fault which integrates NFRPCA modeling and GLRT statistical testing, is shown in Section 4. Next, in Section 5, the GLRT-based NFRPCA test performance is studied using

environmental processes representing the crop model data. At the end, the conclusions are made in Section 6.

I. NONLINEAR FUZZY ROBUST PRINCIPAL COMPONENT ANALYSIS (NFRPCA)

Next, we present the classical principal component analysis.

II.1. Principal Component Analysis (PCA)

Let $X_i \in R^m$ denotes a sample vector of m number of sensors. Also, assume there are *n* samples dedicated to each sensor, a data matrix $X \in R^{nxm}$ is with each row, displaying a sample. Meanwhile, X matrix is scaled to zero mean for covariance-based PCA and at the same time, to unit variance for correlation-based PCA [13]. The *X* matrix can be divided into two matrices: a score matrix S and a loading matrix W through singular value decomposition (SVD):

$$X = SW^T \quad (1)$$

where $S = [s_1 s_2 .. s_m] \in R^{mxm}$ is a transformed variables matrix, $s_i \in R^n$, are the score vectors or principal components, and $W = [w_1 w_2 .. w_3] \in R^{mxm}$ is an orthogonal vectors matrix $w_i \in R^m$ which includes the eigenvectors associated with the covariance matrix of X , i.e., $\sum$, which is given by

$$\sum = \frac{1}{n-1} X^T X = W \Lambda W^T \quad (2)$$

with $W \Lambda W^T = W^T \Lambda W = I_n$ and Λ =diag(λ_1 , λ_2 ,..., λ_m) is a diagonal matrix containing the eigenvalues related to the m PCs, λ_m are simply the eigenvalues of the covariance matrix ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$), and I_n is the identity matrix ([14]). It must be noted at this point that the PCA model yields same number of principal components as the number of original variables (m). Nevertheless, for collinear process variables, a smaller number of principal components (l) are required so that most of the variations in the data are captures. Most of the times, a small subset of the principal components (which correspond to the maximum eigenvalues) might carry the most of the crucial information in a data set, which simplifies the analysis.

The effectiveness of the PCA model depends on the number of principal components (PCs) are to be used for PCA. Selecting an appropriate number of PCs introduces a good performance of PCA in terms of processes monitoring. Several methods for determining the number of PCs have been proposed such as; the Scree plot (Gustafsson 1996), the cumulative percent variance (CPV), the cross validation (Nguyen and Widrow 1990), and the profile likelihood (Willsky et al., 1980). In this study herein, the cumulative percent variance method is utilized to come up with the optimum number of retained principal components. The cumulative percent variance is computed as follows:

$$CPV(l) = \frac{\sum_{i=1}^{l} \lambda_i}{trace(\Sigma)} \times 100 \quad (3)$$

When the number of principal components *l* is determined, then, the data matrix X is shown as the following:

$$X = SW = [\hat{S}\ \tilde{S}][\hat{W}\ \tilde{W}]^T, \quad (4)$$

where $\hat{S} \in R^{nxl}$ and $\tilde{S} \in R^{nx(m-l)}$ are matrices of l retained principal components and the ($m-l$) ignored principal components, respectively, and the matrices $\hat{W} \in R^{mxl}$ and $\tilde{W} \in R^{mx(m-l)}$ are matrices of l retained eigenvectors and the ($m-l$ ignored eigenvectors, respectively. Using Eq. (4), the following can be written:

$$X = \hat{S}\hat{W}^T + \tilde{S}\tilde{W}^T \quad (5)$$

The matrix $\hat{X}$ represents the modeled variation of X based on first l components.

Next, we present the nonlinear fuzzy robust principal component analysis, it is proposed to relate the nonlinear PCA learning rules to energy functions and proposed an objective function with the consideration of noise.

II.2. Nonlinear Fuzzy Robust Principal Component Analysis (NFRPCA)

The PCA might be affected by outliers and a several contributions have been proposed for robustification of PCA (Croux and Haesbroeck 2000; Heo et al., 2009). In addition, outliers are known to influence on the resulting principal component and hence they also have an impact on the modeling as well as fault detection performances. On the other hand, there are many fuzzy approachesin regression analysis which address this issue (Teppola et al., 1999; Xu and Yuille 1995), such that the fuzzy clustering which is an important technique to distinguish between the healthy and faulty structures and identify the structure in the data. To deal with the above problems, we propose to use the objective function *J* proposed in (Gustafsson 1996), where PCA learning rules are related to energy functions with the consideration of outliers:

$$J(W)=\left(\frac{1}{1+(\|X\|^2-\frac{\|W^TX\|^2}{\|W\|^2}/\eta)^{\frac{1}{r-1}}}\right)^r\left(\|X\|^2-\frac{\|W^TX\|^2}{\|W\|^2}\right) \quad (6)$$

The gradient descent rules at every instant of time *t* for estimating the weight *W*(*t*) is given by:

$$W(t+1)=W(t)+\alpha J(W(t)) \quad (7)$$

where, α is the learning rate for the objective function J.
In the case where the size of the covariance matrix is large, it is better to solve the eigenvalue problem by iterative schemes which do not need to compute and store the covariance matrix. To compute iteratively the PCs, the developed iterated robust fuzzy principal component analysis proposes to use the following iterative rules,

A. Fault detection indices

When using PCA in detecting faults, a PCA model is built utilizing fault-free data. The model is used for fault detection through one of the detection indices (the Hotelling's T^2 and Q statistics), which are presented next.

A.1 Hotelling's T^2 statistic

The T^2 statistic is a way of measuring the variation captured in the principal components at various time samples, and it is known as (Dawdle et al., 1982):

$$T^2 = X^T\hat{W}\hat{\Lambda}^{-1}\hat{W}^TX, \quad (8)$$

Where $\hat{\Lambda}^{-1}=diag(\lambda_1,\lambda_2,...,\lambda_l)$, is a diagonal matrix containing the eigen values related to the l retained PCs. For new real-time data, when the value of T^2 statistic exceeds the threshold, T_α^2 calculated as in (Dawdle et al., 1982), a fault is detected.
The threshold number used for the T^2 statistic is computed as (Dawdle et al., 1982):

$$T_\alpha^2=\frac{l(n-1)}{n-1}F_{l,n-l,\alpha}, \quad (9)$$

where α is the level of significance (α usually between 1% and 5%), n is the number of samples in data set, l is the number of retained PCs, and $F_{l,n-l,\alpha}$ is the Fisher F distribution with l and $n-l$ degrees of freedom. These thresholds are computed using faultless data. When the number of observations, n, is high, the T^2 statistic threshold is approximated with a χ^2 distribution with *l* degrees of freedom, i.e., $T_\alpha^2=\chi_{l,\alpha}^2$.

A.2 Q statistic or squared prediction error (SPE)

It is possible to detect new events by computing the squared prediction error SPE or Q of the residuals for a new observation. Q statistic(Tamura and Tsujita 2007; Jackson and Mudholkar 1979), is computed as the sum of squares of the residuals. Also, the Q statistic is a measure of the amount of variation not captured by the PCA model, it is defined as (Tamura and Tsujita 2007):

$$Q=\|\tilde{X}\|^2=\|X-\hat{X}\|^2=\|(I-\hat{W}\hat{W}^T)X\|^2. \quad (10)$$

The monitored system, meanwhile, is accepted to be in normal operation if:

$Q\le Q_\alpha$ (11)The threshold Q_α used for the Q statistic can be computed as (Luukka 2011),

$$Q_\alpha=\varphi_1\left[\frac{h_0c_\alpha\sqrt{2\varphi_2}}{\varphi_1}+1+\frac{\varphi_2h_0(h_0-1)}{\varphi_1^2}\right] \quad (12)$$

where, $\varphi_i=\sum_{j=l+1}^{m}\lambda_j^i,\{i=1,2,3\}$, $h_0=1-\frac{2\varphi_1\varphi_3}{\varphi_2^2}$ and c_α is the value of the normal distribution with α is the level of significance at the instant of an unusual event, when there is a change in the covariance structure of the model, this change is going to be detected by a high value of Q. For new data, the Q statistic is computed and compared to the threshold Q_α (Luukka 2011). This means a fault is detected when the confidence limit is violated. The threshold value is computed on the assumption that the measurements are independent of time and they are multivariate normally distributed. The Q fault detection index is highly sensitive to errors in modeling and the performance of it is dependent on the number of retained PCs, l ,(Zhu and Ghodsi 2006).

II. GENERALIZED LIKELIHOOD RATIO TEST (GLRT)

The faults detection step is done using the residuals computed using PCA. Using the information about the noise distribution of the residuals, a GLR test statistic is formed. To make the decision if a fault is present or not, the test statistic is compared to a threshold from the chi-square distribution.

A. Test Statistic

The GLR test is famous to be a uniformly most powerful test among all invariant tests (shown in Equation (10)). It is basically a hypothesis testing technique which has been

utilized successfully in model-based faults detection (Bulut et al., 2005). Focusing on the following fault detection problem, $Y \in R^n$ is an observation vector formed by one of the two Gaussian distributions: $N(0,\sigma^2 I_n)$ or $N(\theta \neq 0,\sigma^2 I_n)$, where θ is the mean vector (which is the value of the fault) and $\sigma^2 \succ 0$ is the variance (assumed to be known in this problem). The hypothesis test can be shown as:

$$\begin{cases} H_0 = \{Y \sim N(0,\sigma^2 I_n)\} \ (\text{null hypothesis}); \\ H_1 = \{Y \sim N(\theta,\sigma^2 I_n)\} (\text{alternative hypothesis}). \end{cases} \quad (13)$$

Here, the GLR method replaces the unknown parameter, θ, by its maximum likelihood estimate. This estimate is computed by maximizing the generalized likelihood ratio $\mathrm{T}(Y)$ as shown below:

$$\begin{aligned} \mathrm{T}(Y) &= 2 \log \frac{\sup_{\theta \in R^n} f_\theta(Y)}{f_{\theta=0}(Y)} \\ &= 2 \log \left\{ \frac{\sup_{\theta \in R^n} \exp\left(\frac{\|Y - \theta\|_2^2}{f_{\theta=0}(Y)} \right)}{\exp\left(\frac{\|Y\|_2^2}{2\sigma^2} \right)} \right\} \\ &= \frac{1}{2\sigma^2} \left\{ \min_{\theta \in R^n} \|Y - \theta\|_2^2 + \|Y\|_2^2 \right\} \\ &= \frac{1}{2\sigma^2} \left\{ \|Y\|_2^2 \right\} \end{aligned} \quad (14)$$

where $\hat{\theta} = \arg\min \|Y - \theta\|_2^2 = Y$ is the maximum likelihood estimate of θ, the probability density function of Y is $\frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{\|Y-\theta\|_2^2}{2\sigma^2} \right)$, $\|\cdot\|_2$ represents the Euclidean norm. Because the GLR test utilized the ratio of distributions of the faulty and faultless data; for the case of non-Gaussian variables, non-Gaussian distributions are required to be utilized. It must be noted that, in the derivation mentioned above, maximizing the likelihood function is equivalent to maximizing its natural logarithm since the logarithmic function is a monotonic function. At this stage, the GLR test then decides between the hypotheses H_0 and H_1 as follows:

$$\begin{cases} H_0 & \text{if } \mathrm{T}(Y) \prec \mathrm{t}_\alpha \\ H_1 & \text{else.} \end{cases} \quad (15)$$

Since distribution of the decision function $\mathrm{T}(Y)$ under H_0 allows to design a statistical test with a desired false alarm rate, α, where the threshold t_α is chosen to satisfy the following false alarm probability:

$$P_0(\Lambda(Y) \geq \mathrm{t}_\alpha) = \alpha \quad (16)$$

where, $P_0(A)$ represent the probability of an event A when Y is distributed according to the null hypothesis H_0 and α is the desired probability of the false alarm. Since Y is normally distributed, the statistics T is distributed according to the χ^2 law with $(m-l)$ degrees of freedom.

B. Statistic

To select an appropriate thresholds for the test statistics shown above, it is crucial to find their distributions. For that purpose, with the Gaussian noise within, the test statistics will be chi-square distributed variables ([22]). The normalized residual $\overline{R}$ is distributed as

$$\overline{R} \sim N(\theta,\sigma^2 I_n), \quad (17)$$

where $\theta = 0$ under the null hypothesis (15). Then, the test statistic is distributed as the non-central chi-square distribution as shown below:

$$t_\alpha = \frac{1}{\sigma^2} \left\{ \|Y\|_2^2 \right\} \sim \chi_n^2, \quad (18)$$

and the test statistic is distributed through the central chi-square distribution χ_n^2 with degree of freedom n. The threshold is now chosen from the chi-square distribution therefore the fault-free hypothesis is erroneously rejected with only a small probability.

III. FAULT DETECTION USING A GLR-BASED NFRPCA TEST

In this section, a GLR test to detect faults is derived, and its explicit asymptotic statistics computed using PCA. The objective of the GLR-based PCA fault detection technique is to detect the additive fault, θ, with the maximum detection probability for a given false alarm. Here, the fault detection task can be considered as a hypothesis testing problem with consideration of two possible hypotheses: null hypothesis of no change H_0, where measurements vector X, is fault-free, and the change-point alternative hypothesis H_1, where X contains a fault, and thus X is no longer categorized by the fault-free PCA model (4). For new data, the method needs to pick between H_0 and H_1 for the most efficient detection performance. In the absence of a fault, the residual can be calculated as follows,

$$R = X - \hat{X}, \quad (19)$$

while in the presence of an additive fault vector, θ, the residual is computed as,

$$R = X - \hat{X}[+\theta] \quad (20)$$

It is assumed that the residual in Equation (19) is Gaussian. Hence, the fault detection problem consists of detecting the presence of an additive bias vector, θ, in the residual vector, R. The residual vector can be considered as a hypothesis testing problem by focusing on two hypotheses: the null hypothesis H_0, where R is fault-free and the alternate hypothesis H_1, where R contains a fault. The formulation of the hypothesis testing problem can be written as,

$$\begin{cases} H_0 = \{R \sim N(0, \sigma^2 I_n)\} \ (\text{null hypothesis}); \\ H_1 = \{R \sim N(\theta, \sigma^2 I_n)\} (\text{alternative hypothesis}). \end{cases} \quad (21)$$

The algorithm which studies the developed GLR-based PCA fault detection technique is presented in Algorithm 1. The GLR-based PCA is proposed to detect the faults in the residual vector obtained from the PCA model, through which the GLR test is used for each residual vector, R.

Algorithm 1: GLR-based NFRPCA fault detection algorithm.

Input: $N \times m$ data matrix X, Confidence interval α

Output: GLR statistic T, GLR Threshold t_α

- *Data preprocessing step:*

Standardize: computes data's mean and standard deviation, and standardize it;

- *NFRPCA running step:*

Compute the covariance matrix, Σ;

Calculate the eigen values and eigenvectors of Σ and sort the eigen values in decreasing order;

Compute the optimal number of principal components to be used using the *CPV* method;

Compute the sum of approximate and residual matrices;

Testing step:

Standardize the new data;

Generate a residual vector, R, using NFRPCA;

Compute the GLR statistic T for the new data;

Compute the GLR statistic threshold t_α: if $\mathrm{T} \geq t_\alpha$, then declare a fault.

SIMULATION RESULTS ANALYSIS

Next, the crop model that are used to generate data is described.

A. Crop model

The original data were issued from experiments carried out on a silty soil in Belgium, with a wheat crop (Triticum aestivum L., cultivar Julius), during the crop seasons 2008-2009 and 2009-2010. The measurements were the results of 4 repetitions by date, each one of them being performed on a small block (2m times 6m) randomly spread over the field to ensure the measurements independence. A wireless monitoring system (eKo pro series system, Crossbow) completed by a micro-meteorological station was used for measuring continuously soil and climate characteristics. Especially, the measurements of soil water content were performed at 20 and 50 cm depth. The plant characteristics (LAI and biomass) were measured at regular intervals (2 weeks) along the crop seasons, since the middle of February (around Julian day 410) till harvest. Each LAI and biomass measurements were the results of four replicates by date of sampling. The LAI is defined as one half of the total leaf area per unit ground surface area (Jolli?e et al., 2002). Each LAI sample was collected as a 50 cm linear sample (for a total of 2 meters considering four replicates). The stripped leaves were sticked on a paper sheet and digitalized (Chen et al., 1996). The images were segmented using the Meyer and Neto (2008) indices (ExG-ExR) to compute the total green leaf area and the LAI was finally computed as the ratio between this value and the soil reference surface (2 meters times 0.146 meter of inter-row spacing). Each biomass measurement was performed on three adjacent rows of 50 cm (for a total of 6 meters considering the four replicates). The cut samples were dried at laboratory and the total mass was finally weighed. During the season 2008-2009, yields were quite high and close to the optimum of the cultivar. This is mainly explained by the good weather conditions and a sufficient nitrogen nutrition level. The season 2009-2010 was known to induce deep water stresses, and was thus characterized by yield losses. The model for which the methods are tested is Mini-STICS model (Croux and Haesbroeck 2000).

The model equations are presented in (Heo et al., 2009), and the model parameters presented (Teppola et al., 1999). The dynamic equations indicates the way each state variable changes from one day to another as a function of the current values of the state variables, and of the parameters value. Encoding these equations over time allows for eliminating the intermediate values of the state variables and relate the state variables at any time to the explanatory variables on each day. The model structure can be derived from the basic conservation laws, namely material and energy balances (Mourad and Bertrand-Krajewski 2002).

B. Data generation

Indeed, the findings might depend on the details of the model, on the way/quality the data are generated/measured with and on the specific data which was used. To be independent of these consideration, we are generate dynamic data from the crop model. The model is first used to simulate the responses of the 6 state variables: the leaf-area index LAI; LAI, the biomass growth; MASEC, the grain yield MAFRUIT, the volumetric water content of the soil layer1; HUR1, the volumetric water content of the soil layer 2; HUR2, the volumetric water content of the soil layer 3;HUR3 as functions of time of the first recorded climatic

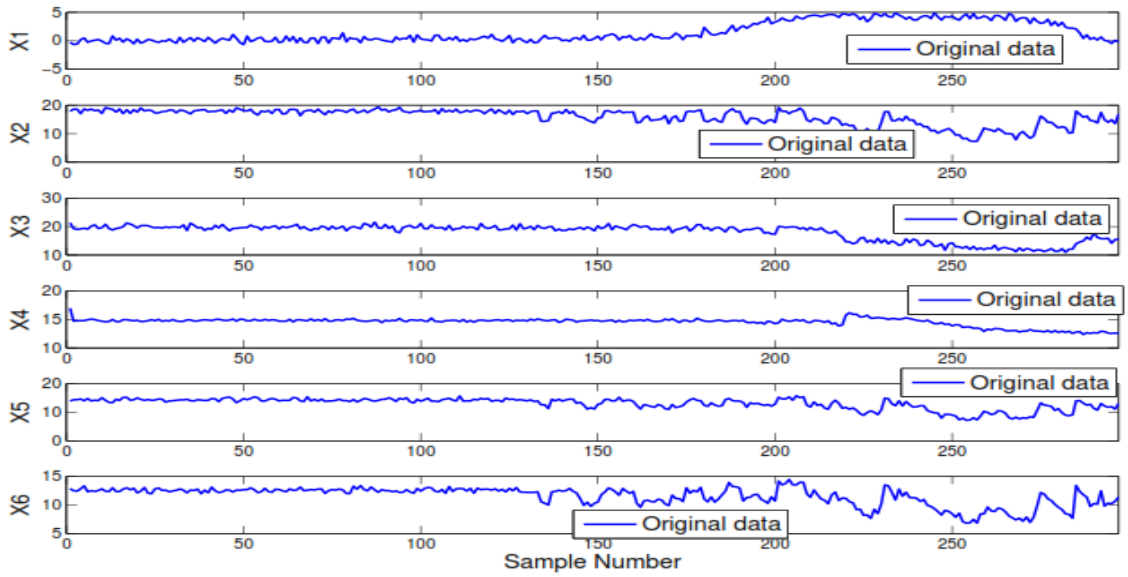

Figure 1. Original data

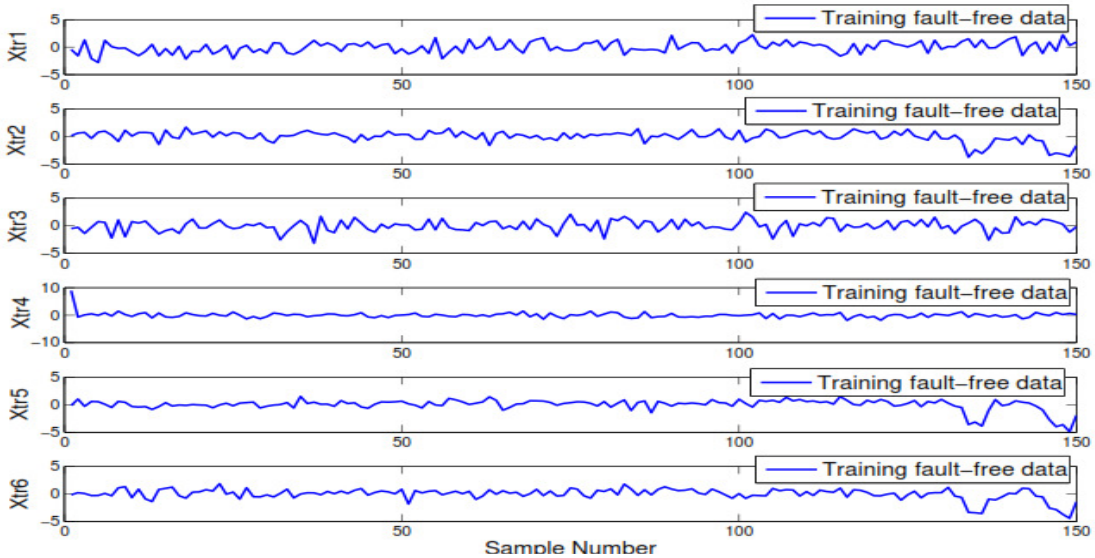

Figure 2. Training fault-free data

variable of the crop season 2008-2009. These simulated states are assumed to be noise free. They are then contaminated with zero mean Gaussian errors, i.e., a measurement noise v). The data set consists of 8 random variables, which are generated using the crop model presented in (Heo et al., 2009, Mansouri et al., 2014). The generated data were arranged as a matrix X having 297 samples and 6 crop model measurements. The responses of the 6 state variables LAI, MASEC, MAFRUIT, HUR1, HUR2 and HUR3, are shown in Figure 1(Mourad and Bertrand-Krajewski 2002).

C. Training of NFRPCA model

As described in Algorithm 1, the NFRPCA -based GLR fault detection method requires constructing a NFRPCA model from fault-free data. Therefore, the fault-free crop model training data described earlier were used to construct a NFRPCA reference model to be used in fault detection. The fault-free crop model data were arranged as a matrix Xtr having 150 rows (samples) and 6 columns (crop model measurements). These data are first scaled (to have zero mean and unit variance), and then are used to construct the NFRPCA model. The responses of the training fault-free data, are shown in Figure 2. The training fault-free data

matrix is used to construct a NFRPCA model. In NFRPCA, most of the crucial variations in the data set are typically captured in the main principal components corresponding to the maximum eigen values as shown in Figure 3. In this study herein, the cumulative percent variance (CPV) method is utilized to find out the optimum number of retained principal components. Utilizing a CPV threshold value of 90%, only the first five principal components of the total variations in the data as displayed in Figure 3 will be retained. A plot of the decision function of the GLR test

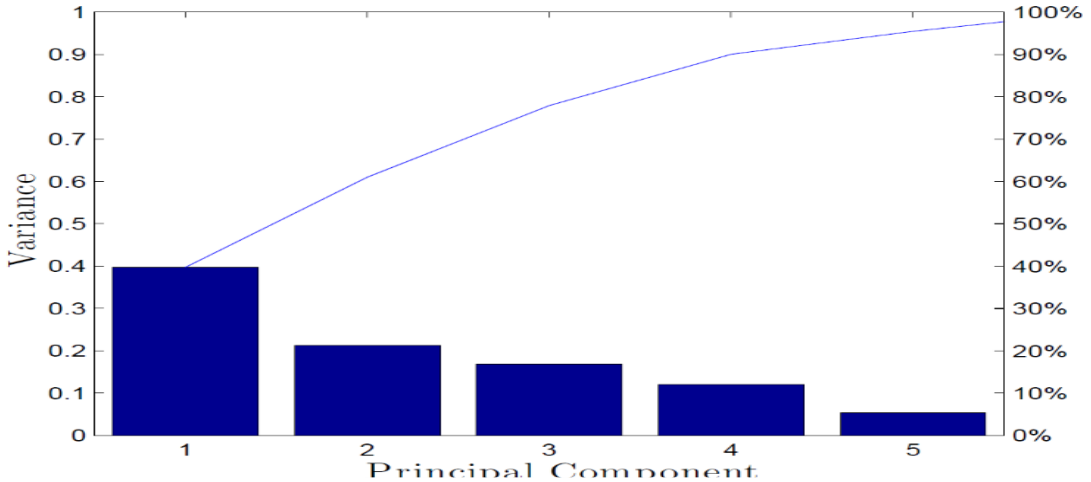

Figure 3. Variance captured by each principal component

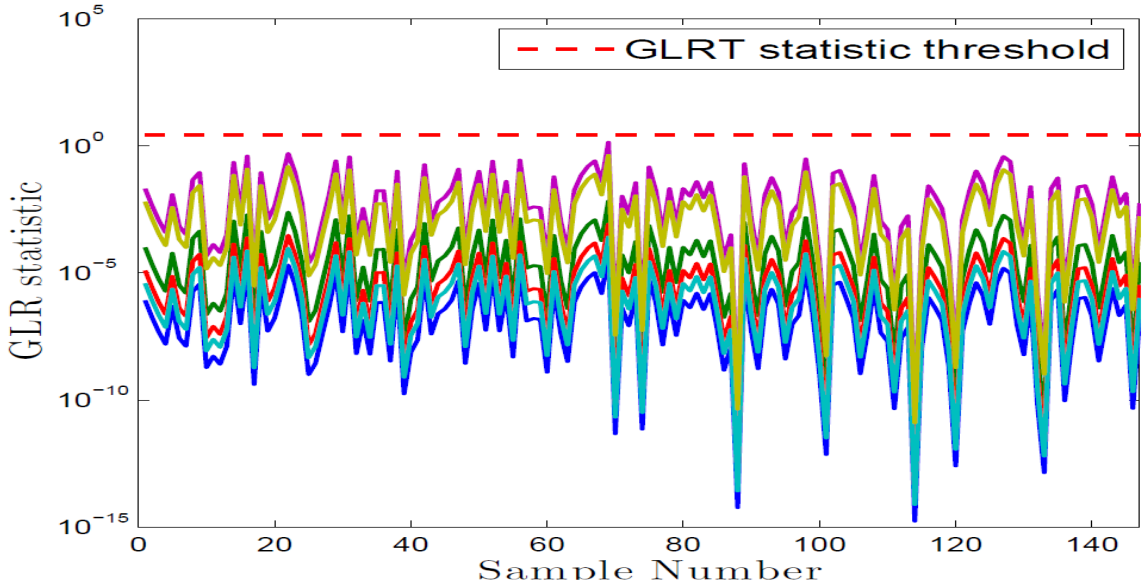

Figure 4. The time evolution of GLR decision function on a semi-logarithmic scale for the fault-free data

(shown in Figure 4) confirms that the process operates under normal conditions, where no faults are present.

D. Fault detection in crop model

The NFRPCA model formed utilizing the fault-free data is deployed in this section to detect possible faults with unseen testing data. Now, the performances of the different fault detection indices will be assessed. To show the abilities of NFRPCA-based GLRT technique in terms of fault detection, we have compared it to the NFRPCA indices T^2 and Q, through three different cases of faults, i) an additive fault (single fault) was introduced in X_1, it consists of a bias of

amplitude equal to 20% of the total variation in X_1, between sample numbers 30 and 80, ii) a double faults were introduced in X_1 and X_2 and iii) multiple faults were introduced in X_1 to X_2. Based on the first three PCs, NFRPCA-based Q, NFRPCA-based T^2 and NFRPCA-based GLRT techniques are used for fault detection (as shown in Figures 6, 7 and 8) in the presence of a single fault in X_1. The results of NFRPCA-based Q statistic is shown in Figure 6, where the dotted line represents the detection threshold Q_α, which is found to be0.6848.

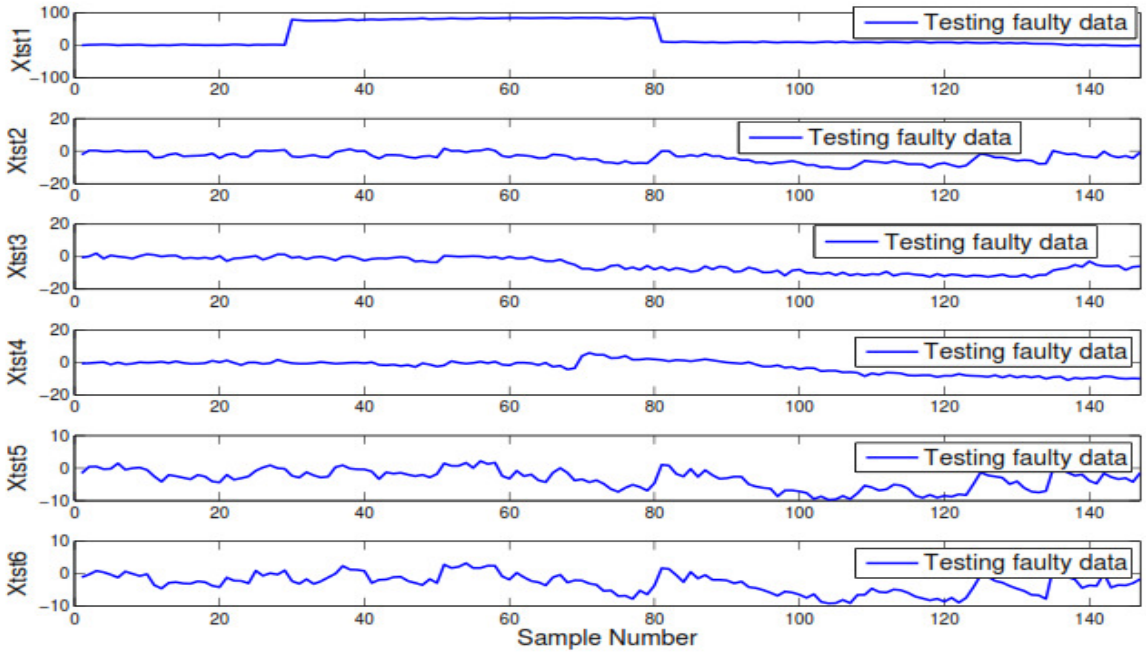

Figure 5. Testing faulty data X_{test}.

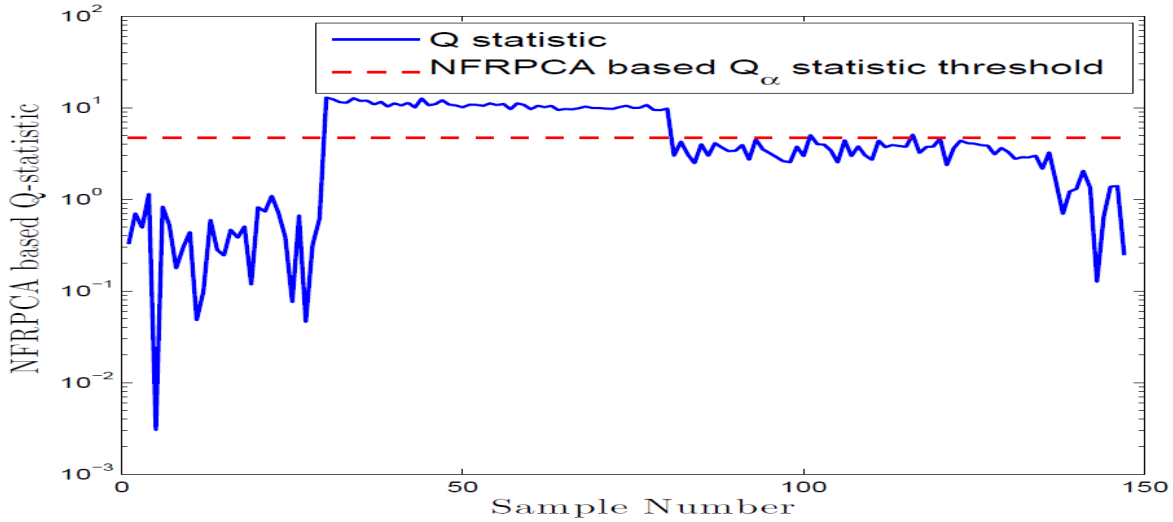

Figure 6. Fault detection using Q statistic in the presence of simple fault.

Figure 7 presents the results of the NFRPCA-based T^2 statistic, where the dotted line represents the detection threshold T^2, T_α^2, which is found to be 9.7. When the NFRPCA-based GLRT is applied using the same fault-free data, the GLRT threshold value is found to be t_α = 553.1 for a false alarm probability of α = 5%. We can show from Figure 5, that, unlike the NFRPCA-based T^2 statistic which results in some missed detections, both NFRPCA-based Q and NFRPCA-based GLRT methods are able to detect the fault effectively (see Figures 6 and 8). The fault is identified at the interval [30 . . . 80] by parallel test of residual subspace (as shown in Figure 5).

Double faults in state variable *X*1 are introduced at the intervals [30 . . . 80] respectively. These faults are represented by a constant bias of amplitude equal 20% of the total variation in X_1. We can show from Figures 9, 10 and 11 the results using NFRPCA-based *Q*, NFRPCA-based T^2 and NFRPCA-based GLRT techniques for faults detection. Figure 9 shows the ability of NFRPCA-based Q, NFRPCA-based T^2 and NFRPCA-based GLRT techniques to detect these additive faults, with some missed detections when using the NFRPCA-based T^2 and Qstatistics (as shown in Figures9 and 10).

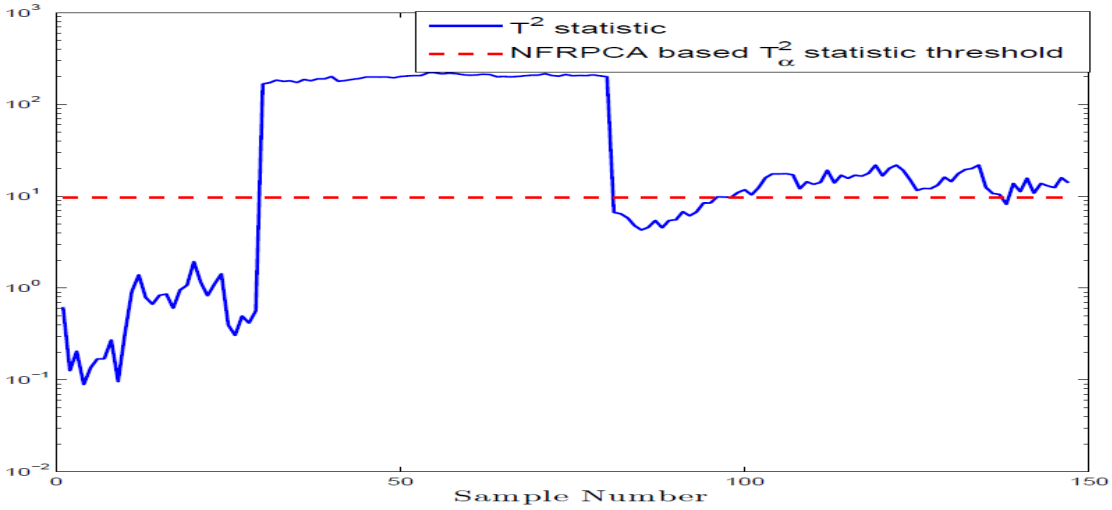

Figure 7. Fault detection using Hotelling's T^2-statistic in the presence of simple fault.

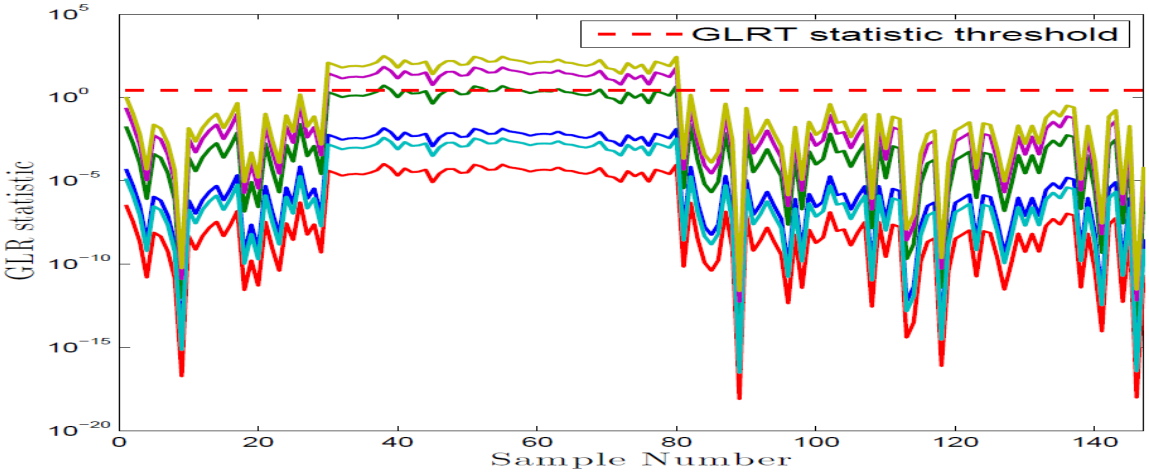

Figure 8. Fault detection using GLR statistic in the presence of simple fault

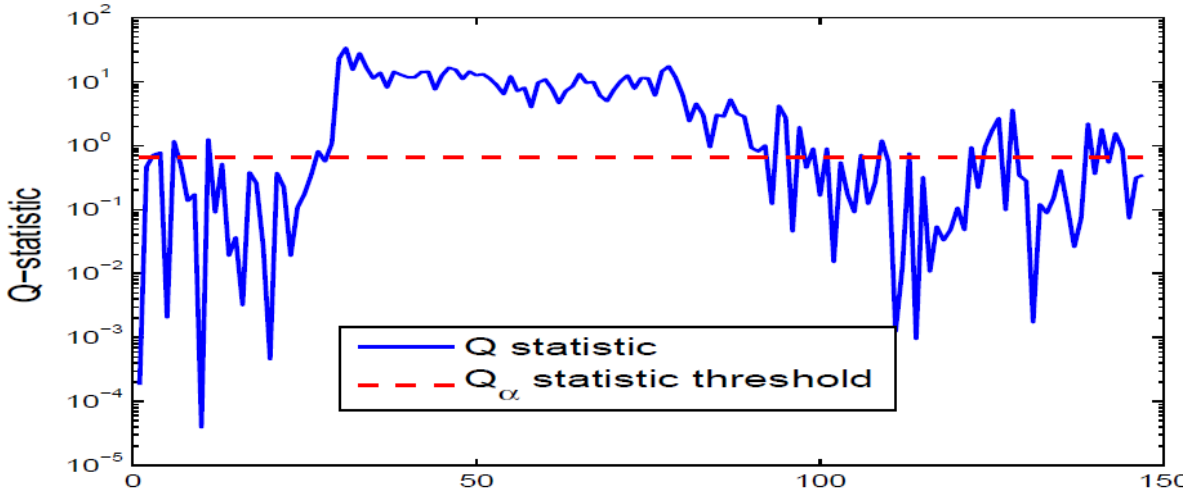

Figure 9. Fault detection using Q-statistic in the presence of double faults.

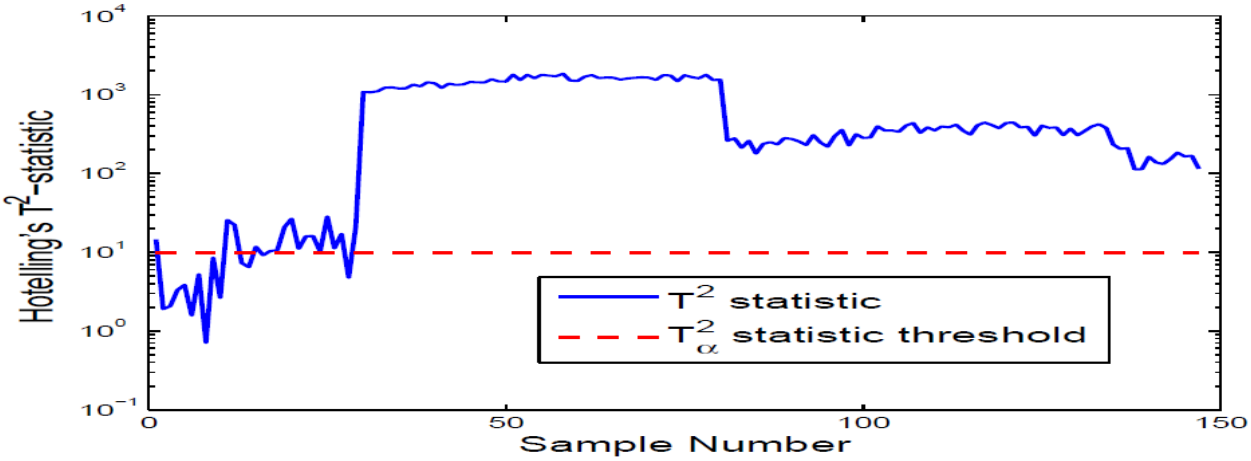

Figure 10. Fault detection using Hotelling's T^2-statistic in the presence of double faults

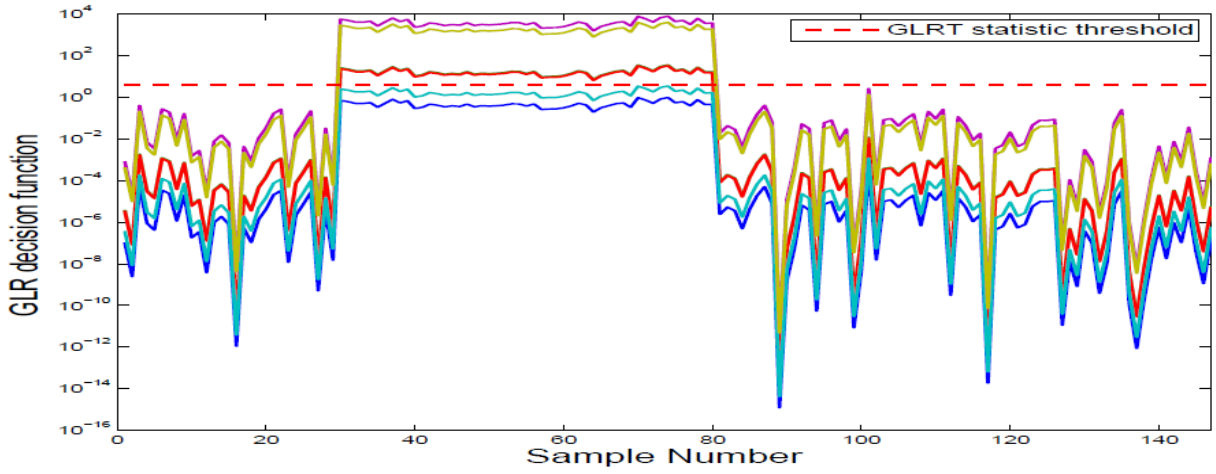

Figure 11. Fault detection using GLR statistic in the presence of double faults.

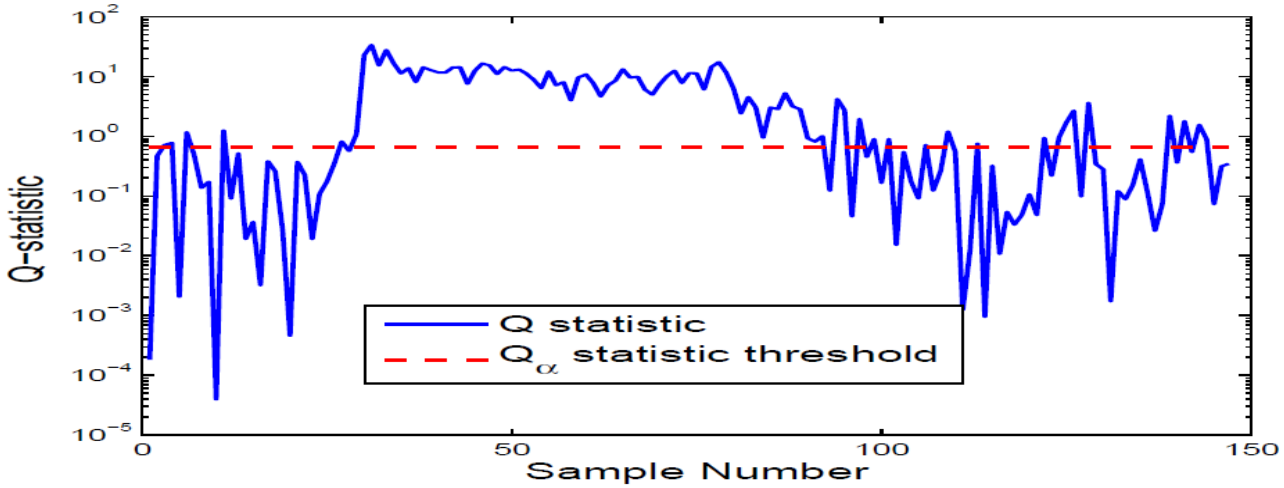

Figure 12. Fault detection using Q-statistic in the presence of triple faults.

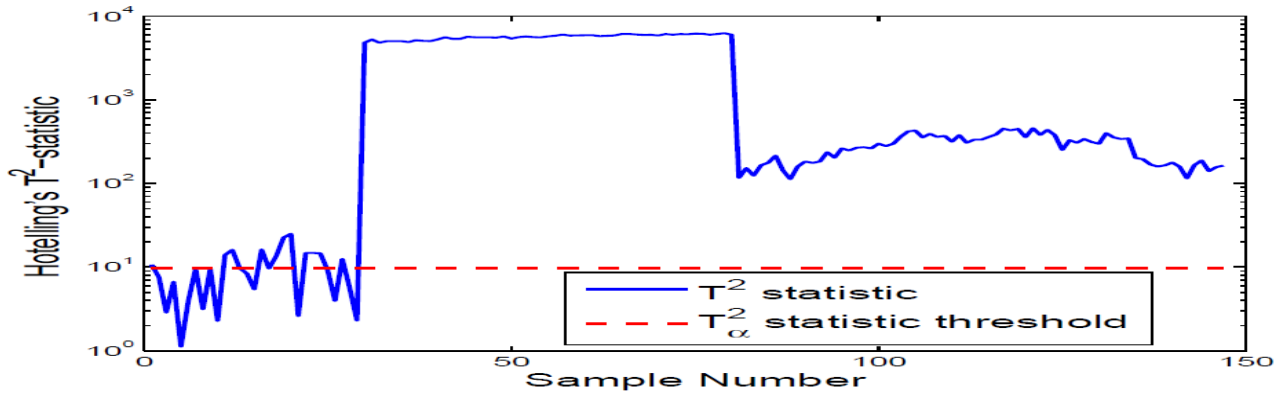

Figure 13. Fault detection using Hotelling's T^2-statistic in the presence of triple faults.

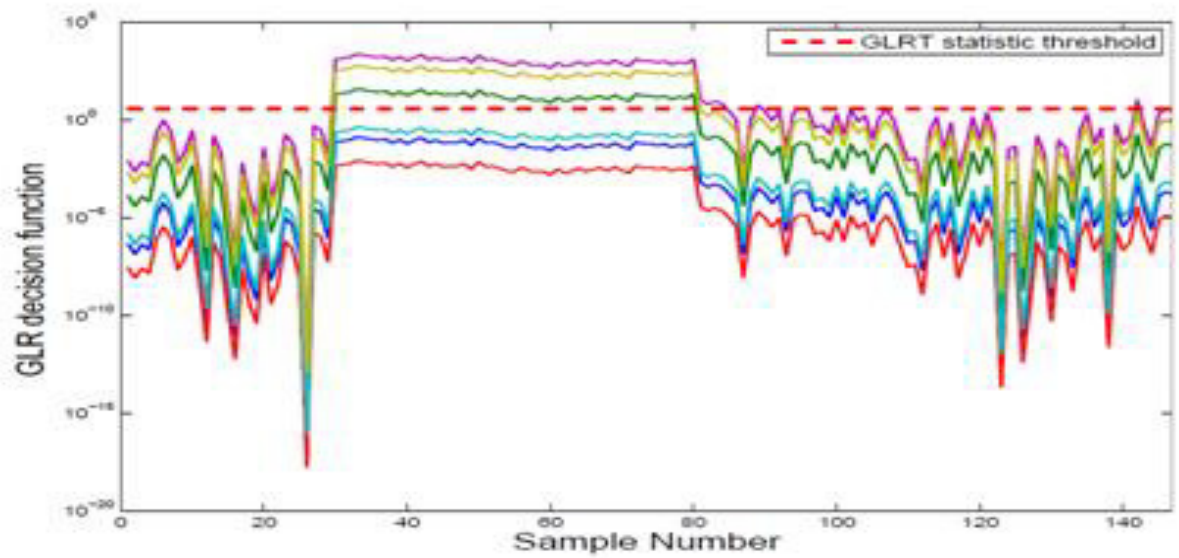

Figure 14. Fault detection using GLR statistic in the presence of triple faults.

The same results are drawn when multiple faults are introduced in X_1 at the intervals [30 . . . 80]. Figures 12, 13 and 14 show that, unlike the NFRPCA-based Q and NFRPCA-based T^2 methods which result in some false alarms(see Figure 12) and missed detections (see Figure 13), the NFRPCA-based GLRT method is able to detect the multiple faults without any false alarms (as shown in Figure 14).

CONCLUSION

In this work, we used nonlinear fuzzy robust principal component analysis (NFRPCA)-based generalized likelihood ratio test (GLRT) for nonlinear fault detection. The fault detection problem was addressed so that the data are first modeled using the NFRPCA method and then the faults are detected using GLRT. The NFRPCA method is investigated here as modeling algorithm in the task of fault detection. The idea is to improve the GLRT performance introducing modeling of the data using the NFRPCA. The NFRPCA method has been proposed to deal with an online are tensions of PCA and provide a good performance over the linear versions. The NFRPCA-based GLRT fault detection performance is assessed and compared to that of the conventional NFRPCA through crop model data. The results demonstrate the effectiveness of the NFRPCA-based GLRT technique over the conventional NFRPCA through its two charts T^2 and Q for detection of single as well as multiple sensor faults.

Acknowledgment

The authors gratefully acknowledge financial support from the Fonds de la Recherche Scientifique - FNRS.

REFERENCES

Benaicha A, Guerfel M, Boughila N, Benothman K (2010). "New pca-based methodology for sensor fault detection and localization," in*MOSIM'10*, Hammamet - Tunisia, May 10-12 2010.

Richard J. Bolton, David J. Hand, David JH (2001). "Unsupervised profiling methods for fraud detection," *Credit Scoring and Credit Control VII*, pp. 235–255, 2001.

Bulut A, Singh AK, Shin P, Fountain T, Jasso H, Yan L, Elgamal A (2005). "Real-time nondestructive structural health monitoring using support vector machines and wavelets," in *Nondestructive Evaluation for Health Monitoring and Diagnostics*. International Society for Optics and Photonics, 2005, pp. 180–189.

Chen J, Bandoni JA, Romagnoli JA (1996). "Robust pca and normal region in multivariate statistical process monitoring," *AIChE journal*, vol. 42, no. 12, pp. 3563–3566, 1996.

Chiang LH, Braatz RD, Russell EL (2001). *Fault detection and diagnosis in industrial systems*. Springer, 2001.

Croux C, Haesbroeck G (2000). "Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies," *Biometrika*, vol. 87, no. 3, pp. 603–618, 2000.

David Z, Marta B (2008). "From large chemical plant data to fault diagnosis integrated to decentralized fault-tolerant control: Pulp mill process application," *Industrial and Engineering Chemistry Research, vol47*, no. 4, pp. p1201–1220, 2008.

Dawdle JR, Willsky A, Gully SW (1982). "Nonlinear generalized likelihood ratio algorithms for maneuver detection and estimation," in *American Control Conference, 1982*. IEEE, 1982, pp. 985–987.

Diana G, Tommasi C (2002). "Cross-validation methods in principal component analysis: A comparison," *Statistical Methods* and *Applications*, vol. 11, no. 1, pp. 71–82, 2002.

Gonzalez F, Dasgupta D, Kozma R (2002). "Combining negative selection and classification techniques for anomaly detection," in *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, vol. 1. IEEE, 2002, pp. 705–710.

Gustafsson F (1996). "The marginalized likelihood ratio test for detecting abrupt changes," *IEEE Transactions on Automatic Control*, vol. 41, no. 1,pp. 66–78, 1996.

Heo G, Gader P, Frigui H (2009). "Rkf-pca: robust kernel fuzzy pca," *Neural networks*, vol. 22, no. 5, pp. 642–650, 2009.

Hotelling H (1993). "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp.417–441, 1933.

Jackson J, Mudholkar G (1979). "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, p.341˝U349, 1979.

Jackson JE, Mudholkar GS (1979). "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21,no. 3, pp. 341–349, 1979.

John GH (1995). "Robust decision trees: Removing outliers from databases." in *KDD*, 1995, pp. 174–179.

Jolliffe I (2002). "Principal component analysis," *second edition, Springer, Berlin*, 2002.

Jonckheere I, Fleck S, Nackaerts K, Muys B, Coppin P, Weiss M, Baret F (2004). "Review of methods for in situ leaf area index determination: Part i. theories, sensors and hemispherical photography," *Agricultural and Forest Meteorology*, vol. 121, no. 1, pp. 19–35, 2004.

Kay SM (1998)."Fundamentals of statistical signal processing: Detection theory, vol. 2," 1998.

Luukka P (2011). "A new nonlinear fuzzy robust pca algorithm and similarity classifier in classification of medical data sets," *International Journal of Fuzzy Systems*, vol. 13, no. 3, pp. 153–162, 2011.

MaGregor J, Kourti T (1995). "Statistical process control of multivariate processes," *Control Engineering Practice*, vol. 3, no. 3, pp. 403–414, 1995.

Mansouri M, Destain MF, Nounou H, Nounou M (2016). "Enhanced monitoring of environmental processes," *International Journal of Environmental Science and Development*, vol. 7, no. 7, p. 525, 2016.

Meyer G, Neto J (2008). "Verification of color vegetation indices for automated crop imaging applications," *Computers and Electronics in Agriculture*, vol. 63, no. 2, pp. 282–293, 2008.

Mourad M, Bertrand-Krajewski JL (2002). "A method for automatic validation of long time series of data in urban hydrology," *Water Science &Technology*, vol. 45, no. 4-5, pp. 263–270, 2002.

Mussardo G (2010). *Statistical field theory*. Oxford Univ. Press, 2010.

Nguyen D, Widrow B (1990). "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*. IEEE, 1990, pp. 21–26.

Oja E (1995). *The nonlinear PCA learning rule and signal separation: Mathematical analysis*. Citeseer, 1995.

Qin S (2003). "Statistical process monitoring: Basics and beyond," *Journal of Chemometrics*, vol. 17, no. 8/9, pp. 480–502, 2003.

Ramaswamy S, Rastogi R, Shim K (2000). "Efficient algorithms for mining outliers from large data sets," in *ACM SIGMOD Record*, vol. 29,no. 2. ACM, 2000, pp. 427–438.

Rousseeuw PJ, Ruts I (1996). "Algorithm as 307: Bivariate location depth," *Applied Statistics*, pp. 516–526, 1996.

Russell EL, Chiang LH, Braatz RD (2000). "Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, no. 1, pp. 81–93, 2000.

Ruts I, Rousseeuw PJ (1996). “Computing depth contours of bivariate point clouds,” *Computational Statistics* &*Data Analysis*, vol. 23, no. 1,pp. 153–168, 1996.

Tamura M, Tsujita S (2007). “A study on the number of principal components and sensitivity of fault detection using pca,” *Computers* and *Chemical Engineering*, vol. 31, no. 9, pp. 1035–1046, 2007.

Tang J, Chen Z, Fu AW-c, Cheung D (2001). “A robust outlier detection scheme for large data sets,” in *In 6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*. Citeseer, 2001.

Teppola P, Mujunen SP, Minkkinen P (1999). “Adaptive fuzzy c-means clustering in process monitoring,” *Chemometrics and intelligent laboratory systems*, vol. 45, no. 1, pp. 23–38, 1999.

Tremblay M, Wallach D (2004). “Comparison of parameter estimation methods for crop models,” *Agronomie*, vol. 24, no. 6-7, pp. 351–365, 2004.

Willsky AS, Chow E, Gershwin S, Greene C, Houpt P, Kurkjian A (1980).“Dynamic model-based techniques for the detection of incidents on freeways,” *Automatic Control, IEEE Transactions on*, vol. 25, no. 3, pp. 347–360, 1980.

Xu L, Yuille AL (1995). “Robust principal component analysis by self-organizing rules based on statistical physics approach,” *Neural Networks,IEEE Transactions on*, vol. 6, no. 1, pp. 131–143, 1995.

Zhu M, Ghodsi A (2006). “Automatic dimensionality selection from the scree plot via the use of profile likelihood,” *Computational Statistics* &*Data Analysis*, vol. 51, pp. 918–930, 2006.

Mansouri, M., Dumont, B., Leemans, V., & Destain, M. F. (2014). Bayesian methods for predicting LAI and soil water content. *Precision agriculture*, 15(2), 184-201.