Reconstruction based fault prognosis for continuous processes

Gang Li, S. Joe Qin, Yindong Ji, Donghua Zhou

1. Introduction

Reliability, maintainability and safety of complex dynamic systems, such as the monitoring system of high speed trains, have attracted much attention in recent years (Labeau, Smidts, & Swaminathan, 2000; Saha, 2003). However, the traditional maintenance and repair strategy, which is so-called ‘run-to-failure’ maintenance, are not adequate, as industrial systems are increasingly called to produce at higher throughput and better quality (Jardine, Lin, & Banjevic, 2006; Kothamasu, Huang, & VerDuin, 2006). Meanwhile, the repairing and inventory cost of some important equipment have become a major cost in the complicated, integrated and interacting systems. To guarantee system safety and decrease maintenance cost, condition-based maintenance and form a basis for further maintenance decision (Jardine et al., 2006).

Li et al. (1999) and Li, Kurfess, and Liang (2000) introduced two methodologies of bearing fault prognosis based on a deterministic and a stochastic defect propagation model with vibration signal analysis. The defect propagation is represented by a mechanism model with time-varying parameters, which offers the best bearing state prediction in the least squares sense. Oppenheimer and Loparo (2002) used a physical relationship between fault severity, machine signatures and remaining life based on crack growth law to estimate remaining useful life (RUL). Qiu, Seth, Liang, and Zhang (2002) related the natural frequency and the acceleration amplitude of a bearing system to its running time and failure lifetime by using system stiffness as a bridge. Thus, the RUL can be predicted online based on vibration measurement.


In the aforementioned research, it is assumed that fault process can be observed directly. However, it is difficult or costly, to observe fault or degradation processes directly in many cases. Wang and Vachtsevanos (2002) used wavelet neural networks as a virtual sensor to produce the fault evaluation information of a bearing crack fault and then applied the fault prognosis with a DWNN (dynamic wavelet neural networks) predictor. Chelidze and Cusumano (2004) considered a hierarchical dynamical system consisting of a directly observable subsystem coupled to...
a slowly time-varying hidden fault process. Both structures were previously known and used to build a new observer for the hidden state, which was then recursively estimated to predict the RUL. Zhang et al. (2005) utilized principal signal features, which were extracted by principal component analysis (PCA), to generate a component health/degradation index using a hidden Markov model (HMM), for further RUL prediction. Xu, Ji, and Zhou (2008) introduced a real-time reliability prediction method for a dynamic system that suffered from a hidden degradation process. This method can also be applied for fault prognosis. Phelps, Willett, Kirubarajan, and Bideau (2007) proposed using failure probabilities of binary random signals to track the system health, which avoided predicting essentially unbounded parameters of a real-valued system. The main idea behind these methods is to monitor the hidden fault process from observations, and then use either mechanical or statistical prediction models to predict the RUL. However, model-based methods need accurate model knowledge (Chelidze & Cusumano, 2004; Xu et al., 2008), which may be difficult and time-consuming for many industrial processes. HMM-based methods assume that system states are discrete and (Zhang et al., 2005), which is more suitable in discrete processes rather than continuous processes.

Approaches based on statistical process monitoring (SPM), which is one of the most active research areas in process control, have been studied actively over the last two decades (Qin, 2003). They are much easier to be applied to real processes than model-based and knowledge-based methods, because of the data-based nature. Instead of a causal model or a mechanistic model, they use empirical correlation model built from normal operating data. Instead of a causal model or a mechanistic model, they use an empirical correlation model built from normal operating data. Instead of a causal model or a mechanistic model, they use an empirical correlation model built from normal operating data.

2. Problem formulation

Let \( x \in \mathbb{R}^m \) denote a sample vector of \( m \) sensors, and \( x^* \) denote the normal sample vector under normal operating conditions. If a fault \( f \) occurs, the sample vector \( x \) can be related with fault \( f \) as follows (Dunia & Qin, 1998b):

\[
x_k = x_k^* + \Xi f_k
\]

where \( \Xi \in \mathbb{R}^{m \times s} \) represents the fault direction matrix related to \( f \), along which the fault develops, and \( f_k \in \mathbb{R}^s \) represents fault magnitude along \( s \) directions of \( \Xi \) at time \( k \). In general, \( s < m \).

The matrix \( \Xi \) is supposed to be known beforehand. For the multidimensional faults, \( \Xi \) is chosen to be an orthonormal matrix. For the unidimensional fault, \( \Xi \) reduces to a vector with unit norm (Dunia & Qin, 1998b). Although there are no state equations here, \( \Xi \) can represent a process fault and a sensor fault (Dunia & Qin, 1998a) as long as the measured variables are affected by the hidden fault process. \( \Xi \) can represent both simple and complex faults which are classified by Yoon and MacGregor (2001). If faulty data are available for a type of fault, fault direction matrix can be extracted from faulty data directly. Examples are given for continuous processes (Valle et al., 2001) and batch processes (Yue & Qin, 2001).

It is a necessary condition for fault prognosis that the hidden fault process is a slowly time-varying autocorrelated process. Note that \( f_k \) changes over time as the actual fault process develops. Thus \( f_k \) is an autocorrelated time series.

3. Fault estimation via reconstruction

As the normal sample vector \( x^* \) cannot be obtained, fault process \( f_k \) cannot be observed directly. To estimate \( f_k \), a data-based process model is required. Effective data-based models exist in the statistical process monitoring area, including principal component analysis (PCA), partial least squares (PLS) and their variants. In this section, a PCA model is used to build a variable correlation model.

3.1. PCA modeling

Let \( x \in \mathbb{R}^m \) denote a sample vector of \( m \) sensors. Assuming there are \( n \) samples for each sensor, a data matrix \( X \in \mathbb{R}^{n \times m} \) consists of \( n \) samples with each row representing a sample and each column representing a sensor. Correlation-based PCA decomposes \( X \) into two parts (referring to Qin, 2003):

\[
X = X + \tilde{X} = TP' + \tilde{X}
\]

where columns of \( X \) are zero-centered and scaled to unit variance. \( \tilde{X} \) represents the modeled variations of \( X \) and \( \tilde{X} \) represents the residual variations. The score and loading matrices, i.e. \( T \in \mathbb{R}^{n \times A} \) and \( P \in \mathbb{R}^{m \times A} \), respectively, can be obtained from eigenvalue decomposition of the sample covariance matrix, with \( A \) the number of significant principal components (typically \( A < m \)) retained such that \( T = XP \) (Valle, Li, & Qin, 1999).

After the decomposition, the variable space is divided into two orthogonal subspaces: the principal components subspace (PCS), \( S_P = \text{span}(P), \) and the residual subspace (RS), \( S_r = \text{span}(1 - PP') \). Then, a sample vector \( X_k \) can be projected onto the PCS and RS, respectively:

\[
x_k = \hat{x}_k + \hat{x}_k
\]

\[
\hat{x}_k = PP'x_k \equiv CX_k \in S_P
\]

\[
\hat{x}_k = (1 - PP')x_k = (1 - CX_k) \in S_r
\]

Note that \( \hat{x}_k^T \hat{x}_k = 0 \), \( dim(S_P) = A \) and \( dim(S_r) = m - A \).
3.2. Fault detection indices

Under the normal operation condition, the variable correlation does not change and the main variation occurs in the PCS. When a fault occurs or an abnormal situation takes place, variable correlation varies, resulting in an increase of the sample projection onto RS. A typical statistic for detecting abnormal conditions is squared prediction error (SPE):

\[ \text{SPE}(x_i) = \|x_i - \hat{x}_i\|^2 = \|I - Cx_i\|^2 \]

(6)

The process is considered normal if

\[ \text{SPE} \leq \delta^2_k \]

(7)

where \( \delta^2_k \) is the upper control limit for SPE with confidence level \( \alpha \).

Jackson and Mudholkar (1979) developed an expression of fault direction matrix, \( \Xi \). One form of the necessary and sufficient condition for complete reconstruction is calculated if and only if the fault can be completely reconstructed.

\[ z_k \quad \text{SPE} \quad \text{eliminating the effect of fault reconstruction is minimized:} \]

\[ n \quad \text{possible. Let} \]

\[ z_k = x_k - \hat{x}_k \]

(8)

The purpose of reconstruction is to find \( \hat{x}_k \) such that the reconstructed SPE is minimized:

\[ \text{SPE}(z_k) = \|z_k\|^2 = \|\hat{x}_k - \Xi \hat{f}_k\|^2 \]

(9)

where \( \hat{z}_k = x_k - \hat{x}_k \).

Dunia and Qin (1998b) studied the reconstructability of multidimensional fault and pointed out that \( \hat{f}_k \) can be uniquely calculated if and only if the fault can be completely reconstructed. One form of the necessary and sufficient condition for complete reconstructability is that \( \hat{z}_k \) has full column rank, which indicates that after \( \hat{z}_k \) is projected onto the RS, its rank does not reduce. It has also been proven that the estimate is unbiased and the variance of estimate error is a constant under this condition.

3.3. Fault estimation via reconstruction

As \( f_k \) cannot be measured directly, fault reconstruction is needed. The objective of fault reconstruction is to estimate normal sample \( x_k^* \) eliminating the effect of fault \( f_k \) as much as possible. Let \( z_k \) denote the reconstruction of \( x_k^* \) from \( x_k \), then \( z_k \) can be calculated as follows:

\[ z_k = x_k - \hat{x}_k \]

(8)

The purpose of reconstruction is to find \( \hat{x}_k \) such that the reconstructed SPE is minimized:

\[ \text{SPE}(z_k) = \|z_k\|^2 = \|\hat{x}_k - \hat{f}_k\|^2 \]

(9)

where \( \hat{z}_k = x_k - \hat{x}_k \).

The optimal solution of this problem gives an estimate of \( \hat{f}_k \) (Qin, 2003):

\[ \hat{f}_k = \hat{z}_k = \hat{z}_k \quad \text{preinverse of} \quad \hat{z}_k \]

(10)

where \( \hat{z}_k \) represents the Moore–Penrose pseudoinverse of \( \hat{z}_k \), and \( \hat{f}_k \) is the estimate of \( f_k \).

There may exist many kinds of faults in a process, denoted by \( f_k \), while the high-frequency part generally represents the noise and/or unknown disturbance. For fault evaluation purposes, it is necessary to extract the lower frequency trend from the initial fault estimation.

Dunia and Qin (1998b) found wide use for noise removal in a variety of fields. This is because they can represent deterministic features in a small number of large coefficients, while stochastic noise affects all wavelet coefficients according to its power spectrum. Thus, the deterministic fault trend (in lower frequency) and stochastic noise or disturbance (in higher frequency) can be separated by an appropriate threshold. Therefore, a wavelet based denoising technology is applied to remove the noise and extract the fault evolution trend in this subsection. However, in most cases, wavelet denoising for real-time signal is actualized via off-line processing. Here, an online wavelet denoising method using a moving window is adopted to denoise each direction of the multi-dimensional signal distribution, one can define the Mahalanobis distance of \( f_k \):

\[ D_k = \hat{f}_k^T R^{-1} \hat{f}_k - \frac{s(n^2-1)}{n(n-s)} F_{s,n-s} \]

(13)

where \( F_{s,n-s} \) is an \( F \) distribution with \( s \) and \( n-s \) degrees of freedom, \( s \) is the dimension of \( f_k \), and \( n \) is the number of normal samples. For a given significance level \( \beta \) the process is considered normal if

\[ D_k \leq D_{\beta} = \frac{s(n^2-1)}{n(n-s)} F_{s,n-s,\beta} \]

(14)

4. Fault prognosis using wavelets and vector AR models

Two assumptions are listed for fault prognosis discussed in this paper:

Assumption 1. The fault process is a slowly time-varying autocorrelated process.

Assumption 2. The fault can be completely reconstructed.

If those conditions are satisfied, fault estimate \( \hat{f}_k \) can be obtained, reflecting how serious the fault is currently. Then, a vector AR model is applied to predict \( \hat{f}_k \) in this section.

4.1. Wavelet based denoising technology

According to Assumption 1, the fault information is located mainly in the low-frequency part of \( \hat{f}_k \), while the high-frequency part generally represents the noise and/or unknown disturbance. For fault evaluation purposes, it is necessary to extract the lower frequency trend from the initial fault estimation.

There are many signal denoising techniques in the time–frequency analysis area. Among these methods, wavelets have found wide use for noise removal in a variety of fields. This is because they can represent deterministic features in a small number of large coefficients, while stochastic noise affects all wavelet coefficients according to its power spectrum. Thus, the deterministic fault trend (in lower frequency) and stochastic noise or disturbance (in higher frequency) can be separated by an appropriate threshold. Therefore, a wavelet based denoising technology is applied to remove the noise and extract the fault evolution trend in this subsection. However, in most cases, wavelet denoising for real-time signal is actualized via off-line processing. Here, an online wavelet denoising method using a moving window is adopted to denoise each direction of the multi-dimensional signal distribution as follows (Xia, Meng, Qian, & Wang, 2007):

\[ \begin{bmatrix} y_{k-n_s-1}(i) \\ \vdots \\ y_k(i) \\ \vdots \\ y_{k+n_s}(i) \end{bmatrix} = T_i^{-1} H_i T_i \begin{bmatrix} \hat{f}_{k-n_s-1}(i) \\ \vdots \\ \hat{f}_k(i) \\ \vdots \\ \hat{f}_{k+n_s}(i) \end{bmatrix} \]

(15)

where \( y_k(i) \) represents fault evolution trend of the ith direction, \( \hat{f}_k(i) \) is the ith element of \( \hat{f}_k \), \( n_s \) is the length of the moving window, \( T_i \) is the wavelet coefficient matrix for the fault in the ith direction, and \( H_i \) is the diagonal filtering matrix for fault in the ith direction.

The low-frequency part of \( \hat{f}_k(i) \) will be extracted efficiently, if the \( H_i \) and \( T_i \) are chosen properly. However, the wavelet basis functions are chosen instead of \( H_i, T_i \) in wavelet based denoising, which will be discussed in the first case study. Besides wavelet basis, there are several factors that may affect the wavelet based
4.2. Fault prognosis using vector AR models

Vector AR (VAR) models have found wide use in describing autoregressed multivariate processes. Thus, a VAR model with $b$ order is used to model the fault process (Deng, 2003):

$$\mathbf{y}_k = \sum_{i=1}^{b} \mathbf{A}_i \mathbf{y}_{k-i} + \mathbf{e}_k$$

where $\mathbf{y}_k \in \mathbb{R}^r$ is the fault trend, $\mathbf{A}_i \in \mathbb{R}^{r \times r}$ $(i = 1, \ldots, b)$ are model parameters, $\mathbf{e}_k \in \mathbb{R}^r$ is the modeling residual, which is often assumed to be zero mean and i.i.d. Let

$$\Theta = [\mathbf{A}_1, \ldots, \mathbf{A}_b]^T$$

$$\varphi_k = \begin{pmatrix} \mathbf{y}_{k-1} \\ \vdots \\ \mathbf{y}_{k-b} \end{pmatrix}$$

The multivariate recursive least squares can be used to estimate $\Theta$ iteratively (Deng, 2003)

$$\begin{cases} \hat{\Theta}_{k+1} = \hat{\Theta}_k + \mathbf{P}_{k+1} [\mathbf{y}_{k+1}^T - \hat{\mathbf{y}}_{k+1}^T \hat{\Theta}_k] \\ \mathbf{P}_{k+1} = \mathbf{P}_k - \frac{\mathbf{P}_k [\mathbf{y}_{k+1}^T \hat{\Theta}_k] [\mathbf{y}_{k+1}^T - \hat{\mathbf{y}}_{k+1}^T \hat{\Theta}_k]^T \mathbf{P}_k}{1 + \hat{\mathbf{y}}_{k+1}^T \mathbf{P}_k \hat{\Theta}_k} \end{cases} \quad (19)$$

where $\hat{\Theta}_0 = 0, \mathbf{P}_0 = \gamma \mathbf{I}$, $\gamma > 0$ is an arbitrary positive number.

The model order $b$ of the VAR model can be determined by Akaike's information criterion (AIC) (De Waele & Broersen, 2003). Then, training set is used to train the model (16) with the algorithm in (19). Once the one step ahead predictor has been obtained, multi-step predictions of $\mathbf{y}_k$ can be calculated in an iterative way:

$$\hat{\mathbf{y}}_{k+p} = \hat{\Theta}^T \varphi_{k+p}$$

where $\hat{\Theta}$ is the estimation from training set, $\varphi_{k+p} = [\mathbf{y}_{k+1}^T, \ldots, \mathbf{y}_{k+p}^T]^T$.

Define the mean square prediction error (MSPE) as follows:

$$\text{MSPE}_p(x) = \frac{1}{L} \sum_{k=1}^{L} (\hat{x}_{k+p} - x_{k+p})^2$$

where $x_k$ is a one dimensional time series with truncated length $L$. $\hat{x}_{k+p}$ represents the $p$ step ahead prediction of $x$. MSPE can be used for evaluating the predictor (20).

Besides VAR models, other modeling approaches can also be used for the prediction of multivariate time series. Choosing a proper predictive model depends on the fault evaluation characteristics. For example, it is better to use a VAR model for linear autocorrelation, a grey model for an exponential trend, and neural networks or support vector machine for a typical nonlinear autocorrelation (Heng, Zhang, Tan, & Mathew, 2009).

Vector AR models are widely used for multivariate time series modeling when the process is autocorrelated linearly. As the fault process is assumed to be slowly varying, it can be treated linear in a short time horizon for prediction. Therefore, it is reasonable to use a vector AR model. Furthermore, the modeling algorithm of vector AR is recursive and adaptable, which has robustness against modeling errors.

4.3. Remaining useful life prediction

When fault process develops, the fault magnitude grows. There exists a control limit for fault prognosis, denoted by $f_{\text{max}}$. When the fault is still tolerable,

$$\|\mathbf{f}_t\| < f_{\text{max}}$$

(22)

The remaining useful life (RUL) at time $k$ is defined using $\mathbf{y}_{k+p}$:

$$\text{RUL}(k) = \min \{ p : \|\mathbf{y}_{k+p}\| \geq f_{\text{max}} \}$$

(23)

The upper limit $f_{\text{max}}$ depends not only on technology factors, such as the fault effect on products and reliability of system, but also on economical factors, such as repair cost and storage of spare parts. Therefore, instead of using statistical distributions, $f_{\text{max}}$ should be manually specified on the basis of process knowledge.

4.4. Summary on the approach

The proposed approach can be summarized in Fig. 1. In the figure, each block represents a procedure of the approach, and the streams describe the dependence relationship among different blocks, e.g. PCA model is built based on the historical normal data. Firstly, the historical normal data are used to build a PCA model (‘PCA’ block), which describes the correlations among all measured variables under normal operation condition. If some kinds of faulty data are collected beforehand, the residual fault directions can be extracted for these faults (‘Residual fault direction’ block). Given a real-time measurement with a known fault, the fault can be detected based on the prebuilt PCA model (‘Fault detection’ block). The fault directions for known types are combined to identify the fault type (‘Fault identification’ block) and further estimate the fault magnitude (‘Fault estimation’ block). Subsequently, the multivariate fault estimation is processed with noise removal (‘Denoise’ block) and then predicted by a vector AR model (‘Fault prediction’ block). Lastly, the RUL can be calculated (‘RUL prediction’ block).

5. Case studies

In this section, two cases studies on CSTR and TEP, respectively, are considered to show the effectiveness of the proposed multivariate fault prognosis approach. The new fault detection index and how to choose the parameters in denoising procedure are also considered.

5.1. Case study on CSTR

Firstly, a case study on continuous stirred tank reactor (CSTR) (Zhou & Ye, 2000, p. 312) with feedback control is used to show the application of fault prognosis in detail.

5.1.1. Process description

The CSTR process can be described by the following group of differential equations:

$$\frac{dC_A}{dt} = \frac{q}{V} (C_{\text{in}} - C_A) - k_0 \exp\left( - \frac{E}{RT} \right) C_A + v_1$$

(24a)

$$\frac{dT}{dt} = \frac{q}{\rho C_p} \Delta H \frac{k_0 \exp\left( - \frac{E}{RT} \right)}{P C_p} (T - T_c) + v_2$$

(24b)

where $C_A$ is the outlet concentration, $T$ is the reaction temperature, $T_c$ is the temperature of cooling water, $q$ is the input Fluent velocity of reactant, $C_{\text{in}}$ is the input reactant concentration, $T_f$ is the input reactant temperature, and $v_1, v_2$ are independent system noise process, where $v_i(k) \sim N(0, \sigma_i^2)$. Other variables are the
simulation parameters, which are constant during the process. In the simulation, $C_N, T$ are the controlled objective with nominal values, and $C_k, q$ are chosen as controllable variables with feedback from control errors. It is assumed that there is a slowly increasing fault process in $C_N$, and $C_k, T$ are not measurable. The observed process variables affected by the fault process are $C_p, T$ and $T_c$.

Measurement noise is added to measurable variables: $x_{\text{meas}} = x(k) + \sigma(k)$, where $x_{\text{meas}}$ means the measurement and $\sigma(k) \sim N(0, \sigma^2)$. The negative feedback inputs are added to $[q, T]^T$ with the transfer function form of PID controller as $K_p (q_k + T + T_1) s$, where $\pi = (C_q, C_p, T) - \pi^*$ is the residual vector between measurements and nominal values. All the system parameters and conditions are listed in Table 1.

### Table 1
Parameters and conditions in CSTR simulation.

| Simulation parameters | $\Delta H = 1783.821$ J/mol, $\rho = 1000$ g/L, $E[R = 5360$ K, $V = 1001$, UA = 1.19501/(min.K), $k_0 = \exp(13.4)$ min$^{-1}$, $C_p = 0.239$ J/(g.K) |
| Initial conditions | $q = 1001$(min), $T_c = 419$ K, $T_f = 400$ K, $C_N = 1$ mol/L |
| Nominal values | $C_p = 0.2$ mol/L, $T_f = 446$ K |
| Controller information | $K_p = 1.7$, $T_q = 0.1$, $T_1 = 10$, $K_c = [5, 1, 2]$ |
| Noise conditions | $\sigma^2_{x_1} = 0.01$, $\sigma^2_{x_2} = 0.01$, $\sigma^2(T) = 0.005$, $\sigma^2(C_p) = 1e^{-5}$, $\sigma^2(q) = \sigma^2(T) = 1e^{-6}$ |

5.1.2. Fault detection using SPE and Df

After mean-centering all variables and scaling all variables into unit variance, a PCA model is built with 1000 normal observations. Based on the cumulative percent variance (CPV) criterion (Valle et al., 1999), three principal components are kept, i.e. $A = 3$. One thousand faulty observations under the condition $C_N = 1.2$ mol/L are used to extract the residual fault direction matrix $\Xi$ according to Valle et al. (2001). As there is only one freedom in subspace $S_x$, the fault dimension is one, i.e. $s = 1$. And the reduced fault direction vector (i.e. $\Xi$) is listed below:

$\Xi = [-0.01895, -0.0048, -0.7052, 0.7088]^T$

A set of faulty data consisting of 1000 samples with the following hidden fault process are used for fault prognosis:

$$
\begin{cases}
C_N(k) = 1, & \text{mol/L, } k < 300 \\
C_N(k) = 1 + (k - 300) / 1500, & \text{mol/L, } k > = 300
\end{cases}
$$

(25)

Measured process variables affected by the hidden fault process, Eq. (25), are shown plotted in Fig. 2. Using PCA based methods, the fault can be detected by SPE index. Given fault direction $\Xi$, the fault detection can also be performed with $D_f$.

Fig. 3 shows the fault detection results using SPE and $D_f$, respectively. The control limit for SPE index is given by Jackson and Mudholkar (1979), while that for $D_f$ is calculated by (14).
From Fig. 3, the detection times given by them are 455 and 431, respectively, which indicates that $D_f$ may detect the fault process more sensitively than the SPE index. This is because more information of fault, i.e. the fault direction, is used in $D_f$. Furthermore, the detection index $D_f$ is in the form of the function of $f_k$, which integrates the fault detection and prognosis together. To sum up, when the fault type can be predetermined, and the fault direction is known beforehand, index $D_f$ is more preferred. Otherwise, SPE index is preferred, and a diagnosis procedure is required before performing fault prognosis.

There is a delay of about 150 min for detection, as the fault process starts at 300 min. This is because the CSTR process is a dynamic process, and the fault process varies very slowly. When the fault is just introduced, the fault effect on process is too small to change SPE or $D_f$. When the fault grows larger, the effect can be detected and then predicted.

5.1.3. Reconstruction based fault estimation

The fault vector is estimated based on reconstruction, as shown in Fig. 4. When the fault is detected by SPE or $D_f$, the fault prognosis starts. The prognosis control limit $f_{max}$, which is set using expert knowledge, determines when to end the prognosis. In this simulation, $f_{max}$ is chosen as 1.45, to facilitate the presentation of RUL prediction.

5.1.4. Wavelet based denoising

The denoising procedure is implemented to the fault estimation after the fault is detected, as shown in Fig. 4. In wavelet based online denoising, the discrete wavelet transform is performed inside moving window with length $n_w$, which is also called Mallat algorithm. The data are first decomposed into $n_d$ layers, computing the scaling coefficients and wavelet coefficients for each layer. Then, wavelets coefficients are shrunk with the soft shrinkage of universal threshold. Lastly, the signal is reconstructed with processed wavelet coefficients and scaling coefficients.

The main parameters in this algorithm are $n_w$, $n_d$ and wavelet basis function. In general, there are several wavelet basis that are suitable for denoising, such as Haar wavelets, Daubechies wavelets ($dbN$) and Symlet wavelets ($symN$). Selecting the proper wavelet basis depends on the noise property. Further, according to application condition of wavelet multi-resolution algorithm, $n_w$ is preferred to have the form $2^L$ ($L > n_d$). In this subsection, Haar wavelet function is used, and $n_d=4, n_w=64$. Parameter selection of wavelet basis, $n_d$ and $n_w$ will be studied in the end of this simulation.

5.1.5. Fault prognosis using VAR models

A model order of 3 was determined using AIC. An autoregressive model with order 3 is built from about 300 samples of $y_k$, while about 200 other samples of $y_k$ are used for prognosis. A one-step predictor is identified by the MRLS algorithm:

$$\hat{y}_k = 1.438y_{k-1} - 0.143y_{k-2} - 0.295y_{k-3}$$  \hspace{1cm} (26)$$

The prediction error for training and testing samples is plotted in Fig. 5. The MSPE of testing data using this model is $1.37e^{-5}$. Using one step ahead predictor (26) to predict $y_{k+p}$ iteratively, the RUL is predicted online. Fig. 6 depicts the RUL prediction when the fault grows around the prognosis control limit. Dashed line represents the real RUL, which is calculated by subtracting current time from the time when the fault reaches the prognosis control limit. The maximum of prediction steps is set to 50, which should be determined by an expert in practice. When the fault grows nearly to the control limit, the predicted RUL drops under 50 steps, which gives an early warning. From Fig. 6, RUL prediction is
effective, which helps to prepare the spares beforehand, and improve the system safety significantly.

5.1.6. Parameter selection in denoising procedure

In this subsection, the parameter selection in wavelet based denoising is studied based on MSPE. Table 2 lists the MSPE of $y_k$ using wavelet based denoising with different parameters, where model order means the order of VAR model, time cost means how long the denoising procedure takes. From Table 2, it is observed that when the decomposition depth $N$ increases, MSPE decreases and the time cost raises. It is also seen that window depth $n_w$ has a limited impact on the ultimate VAR modeling. Different kinds of wavelet basis are compared, and haar wavelet is more proper in this simulation.

5.2. Case study on TEP

In this subsection, the effectiveness of the proposed methods on multidimensional faults is investigated by application to the Tennessee Eastman process (TEP).

5.2.1. Process description

TEP was created by the Eastman Chemical Company to provide a realistic industrial process for evaluating process control and monitoring methods (Downs & Vogel, 1993). The detailed description of the process can be found in Chiang, Russell, and Braatz (2001). TEP has been widely used as a benchmark process for evaluating the process monitoring methods such as PCA, PLS, and Fisher discriminant analysis (FDA) (Chiang, Russell, & Braatz, 2000; Lee, Han, & Yoon, 2004). TEP contains two blocks of variables: the XMV block of 12 manipulated variables and XMEAS block of 41 measured variables. Process measurements are sampled with interval of 3 min. Nineteen composition measurements are sampled with time delays that vary from 6 to 15 min, which are not included in process variables. There are 15 known faults in TEP, where fault 13 is a slow drift in the reaction kinetics, which is suitable for fault prognosis.

5.2.2. Fault subspace extraction and estimation

In this study, 22 process measurements and 11 manipulated variables, i.e. XMEAS(1–22) and XMV(1–11), are chosen as $X$. Firstly, a PCA model is built from 480 normal samples. The samples are centered to zero mean and scaled to unit variance. Based on the cumulative percent variance captured by PCA model, the principal components number is selected as 23. Then, the fault subspace is extracted using 480 faulty samples under fault 13. According to Valle et al. (2001), the dimension of fault vector is 3. The reconstructed SPE data for historical faulty samples with different number of fault dimension are plotted in Fig. 7.
observed that the reconstructed SPE data with 3 dimension fault directions drop below the fault detection limit. The reduced fault direction matrix \( \mathbf{N} \in \mathbb{R}^{33/2} \).

Following that, a set of 960 testing samples under fault 13 is used for prognosis. In the simulation, the fault is introduced in the 160th sample. The fault is detected at sample 202 and 203 using SPE and \( D_f \), respectively, as shown in Fig. 8. There is a delay of about 40 samples, for the similar reason to the CSTR case. After the fault is detected, the fault vector \( f_k \) is estimated based on fault reconstruction. The magnitude and each dimension of \( f_k \) are plotted in Fig. 9.

5.2.3. Fault prediction using wavelets and VAR models

A VAR model is used to predict fault vector \( f_k \in \mathbb{R}^3 \) directly. The model order \( b \) is one according to AIC. The estimated model parameter is

\[
A_1 = \begin{bmatrix}
0.9879 & 0.0150 & -0.0483 \\
0.0168 & 0.9170 & 0.0265 \\
0.0447 & -0.0791 & 1.0042
\end{bmatrix}
\]

Fig. 10 describes the one step ahead and 10 steps ahead prediction result based on the direct prognosis. To reduce the prediction error, a wavelet denoising procedure is adopted. Similar to the CSTR case study, the Haar wavelet is used, and the decomposition layers is set to three, \( n_w = 64 \). Then a VAR model with \( p=2 \) is used for prognosis. The estimated model parameter is

\[
|A_1, A_2| = \begin{bmatrix}
1.7849 & -0.0041 & 0.1016 & -0.7956 & 0.0210 & -0.1126 \\
0.1849 & 1.7159 & -0.0877 & -0.1780 & -0.7470 & 0.1039 \\
-0.0502 & -0.1870 & 1.8959 & 0.0542 & 0.1833 & -0.9001
\end{bmatrix}
\]
Table 3
Mean square prediction error in different cases.

<table>
<thead>
<tr>
<th>MSPE</th>
<th>1st dim</th>
<th>2nd dim</th>
<th>3rd dim</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSPE of $\bar{y}_k$</td>
<td>0.3359</td>
<td>0.2914</td>
<td>0.1645</td>
</tr>
<tr>
<td>MSPE of $\bar{f}_k$</td>
<td>5.3513</td>
<td>4.4192</td>
<td>5.5680</td>
</tr>
<tr>
<td>MSPE of $y_k$</td>
<td>0.0750</td>
<td>0.0934</td>
<td>0.0155</td>
</tr>
<tr>
<td>MSPE of $y_k$</td>
<td>3.0318</td>
<td>2.8529</td>
<td>0.7090</td>
</tr>
</tbody>
</table>

Fig. 11 describes the one step ahead and 10 steps ahead prediction of the denoised fault trend $y_k$. The prediction error by the direct prognosis and denoising based prognosis is listed in Table 3. It is observed that wavelet denoising can reduce the prediction error efficiently, which is significant for multi-step prediction.

6. Conclusions

This paper considers a multivariate fault prognosis method for continuous processes with a hidden fault process. The method uses the fault description widely accepted in statistical process monitoring, which can represent many types of faults. The assumptions of fault reconstructability are given for multidimensional faults. Fault magnitude can be estimated based on reconstruction and a vector AR model with wavelet based denoising is used. Given the fault direction, a new fault index is proposed to integrate detection and prognosis, which is observed to have the same sensitiveness to the fault as the SPE index. The case studies on a CSTR and the Tennessee Eastman process demonstrate the effectiveness of the proposed approaches. Parameters in denoising procedure may affect the prediction model, which should be chosen properly. The concept of remaining useful life prediction is shown clearly in the paper.

Acknowledgements

This work was supported by National 973 project under Grants 2010CB713800 and 2009CB20602 and the NSFC under Grants 60721003, 60736026 and 60931160440. S. Joe Qin acknowledges the financial support from the Changjiang Professorship by the Ministry of Education of PR China.

References


