A Simple Design Strategy for Fault Monitoring Systems

Chuei-Tin Chang, Kai-Nan Mah, and Chii-Shiang Tsai
Dept. of Chemical Engineering, National Cheng Kung University, Tainan, Taiwan 70101, ROC

A simple strategy is developed for the design of model-based fault-monitoring systems. Several new features allow this approach suitable for practical applications. The core of the suggested monitoring system is a parallel parameter estimation method designed to reduce the chance of bias. Two important criteria, diagnostic observability and diagnostic resolution, are adopted to evaluate the performances of alternative designs. By following a systematic procedure, it is possible to determine the best sensor locations in a complex chemical process based only on the structural informations of the system model. After selecting the most appropriate measurement variables, diagnostic resolution can be enhanced further by exploiting additional insights gained from analysis of the physical meanings of the model parameters and by comparing the fault patterns in diagnosis. Several effective on-line implementation techniques are also proposed to minimize computation for the estimation algorithm. The correctness and efficiency of the suggested design strategy are demonstrated by the simulation results in examples.

Introduction

Due to the frequency of serious accidents in chemical industries during recent years, the importance of fault monitoring systems in operating complex modern plants has become apparent. Although numerous approaches have been suggested in the literature, few can be actually implemented in commercial chemical processing units. Thus, the design of fault monitoring systems is not only a challenging research topic, but also a matter of practical significance.

Most recent studies in this area are concerned with the developments of computer-aided diagnosis systems for assisting operators to respond to emergency situations effectively (Davis et al., 1987; Lamb et al., 1987; Petti et al., 1990). Various techniques have been suggested, for example, the signed directed graph (Kramer and Palowitch, 1987; Chang and Yu, 1990; Yu and Lee, 1991), parameter estimation (Isermann, 1984), the expert system (Rich and Venkatasubramanian, 1987; Petti et al., 1990) and neural networks (Venkatasubramanian and Chan, 1989; Watanabe et al., 1989; Venkatasubramanian et al., 1990; Hoskins et al., 1991). In the research presented here, model-based analysis was adopted for the development of fault detection and diagnosis methods. More specifically, detection of changes in states and/or parameters of the mathematical model was achieved by implementing system identification techniques. Diagnosis of the fault origins, for example, equipment malfunctions and external disturbances, were accomplished on the basis of the physical interpretations of these changes and/or structural analysis of the model. As a direct result of this approach, the candidate faults were limited in this research to those that can be associated with parameter variations. In other words, sensor failures and failures that can change the structure of system models were not considered in our analysis.

There are a large number of related works published in the literature (Willsky, 1976; Isermann, 1984; Ljung, 1987). Among various different estimation techniques adopted in the past, the extended Kalman filter (EKF) is clearly one of the most popular methods, see Park and Himmelblau (1983), Watanabe and Himmelblau (1983a,b, 1984), Dalle Molle and Himmelblau (1987), and Li and Olson (1991). In essence, EKFs of one form or another were employed to estimate both the states and parameters of chemical engineering systems. Causes of abnormal system behaviors were then identified accordingly. Although the effectiveness of EKFs has been adequately demonstrated in these published studies, it is a well-known

Correspondence concerning this article should be addressed to Chuei-Tin Chang.

fact that the use of EKF: requires substantial computation time, exhibits slow convergent rate and does not always produce unbiased estimates (Watanabe and Himmelblau, 1984). From the standpoint of fault diagnosis, the most serious drawback of the extended Kalman filter is its inability to guarantee unbiased estimates. Obviously, incorrect information about the system parameters and/or states can mislead diagnosis. To overcome this problem, modifications of the traditional algorithm were introduced in this research. In particular, several smaller EKFs were used in parallel to reduce the chance of generating biased estimates. On the basis of this approach, simple design procedures were then developed to optimize the performance of the fault monitoring system for any given process.

Due to financial limitations, not all of the state variables in a chemical plant can be measured on-line. Thus the problem of selecting the best sensor locations is a critical issue in the design of fault monitoring systems. Two selection criteria were adopted in this study: diagnostic observability and diagnostic resolution. By *observability*, we mean that, in at least one of the parallel EKFs, the parameter corresponding to the actual fault can be estimated correctly. To satisfy the observability condition, a systematic procedure has been developed in this work for choosing the minimum number of variables to be measured on-line. The term *diagnostic resolution* usually refers to a measure of uniqueness in identifying the correct fault origin among all possible candidates (Tsuge et al., 1985; Kramer and Polawitch, 1987). Based on a qualitative analysis of the characteristics of EKF, a simple method has been developed for synthesizing fault monitoring systems with high resolution. Under the condition that the total number of measurement points is fixed, it is possible to determine the most appropriate distribution of sensor locations based only on the structural information of the system model. Since there is no need for quantitative calculations, the best set of measurements can be chosen quickly even for a large and complex process. Finally, several on-line implementation strategies were proposed to reduce the computation load of implementing parallel EKFs in large industrial-scale systems. The feasibility and effectiveness of all the techniques suggested in this research were confirmed by numerical simulation studies and will be elaborated in various examples.

Parallel Parameter Estimation Method

In developing diagnostic procedures, the single-fault assumption has often been adopted, for example, Kramer and Palowitch (1987). Our work is not an exception. This is mainly due to the rationale that the probability of simultaneous occurrence of two or more independent faults is extremely low and therefore can be considered negligible. As mentioned before, the extended Kalman filter is used in this study as a tool for fault detection and diagnosis. Since there are many possible events which could cause the observed system symptoms, the number of augment states in a conventional EKF is usually large. Thus if implemented without modifications, biased estimates may be generated simply because more than one equally satisfactory way of fitting the measurement data exist.

If the parameter associated with the fault origin is known, then there is no need for augmenting all the parameters to the state vector in the corresponding EKF. In this filter, only the parameter associated with the anticipated malfunction is allowed to deviate from its normal value and all other parameters remain unchanged. Thus, there is no chance for faults other than the assumed one to be detected and, since the assumed model is capable of describing the actual system behavior, the possibility of biased estimation can be almost eliminated.

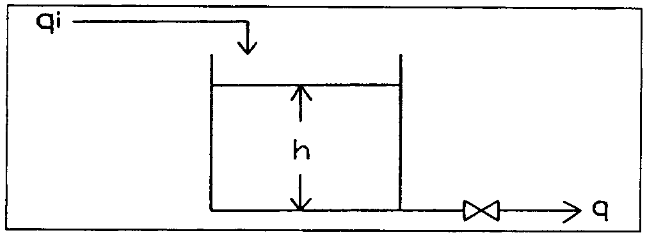

Figure 1. The simplified process flow diagram of a one-tank system.

However, obtaining this information about the deviant parameter is exactly one of the purposes of fault diagnosis. Thus, it becomes necessary to implement several smaller EKFs in parallel, each of them corresponding to one of the parameters, in order to include all the possibilities, for example, Willsky (1976). This approach, although simplistic, is quite effective and will be referred to as the *parallel parameter estimation method* throughout this article. The following example demonstrates the advantage of using this modified approach:

Example 1. Let us consider the tank system in Figure 1. Assume that three faults, a leak in tank, partial blockage in the outlet pipeline and a change in the inlet flow rate, are the most likely to occur in this system. The following is the model for a conventional EKF (Denn, 1980):

$$A\frac{dh}{dt}=q_i-q-c_l\sqrt{h} \tag{1a}$$

$$\frac{dq}{dt}=\frac{\pi d^2}{4\rho l}\left[\rho g h-(f+f_v+f_p)\frac{8l\rho q|q|}{\pi^2 d^5}\right] \tag{1b}$$

where A represents the cross-sectional area of the tank, h is the height of the liquid level, t is the time, q denotes the outlet flow rate, d and l represent the diameter and length of the outlet pipeline, f is the friction factor for the flow in the outlet pipeline, which can be calculated by the correlation equation suggested by Churchill (1977), and f_v is used to account for the additional friction caused by the valve on the outlet pipeline, which was assumed to remain constant at 0.02. To detect the occurrence of any one of the candidate faults, three additional differential equations must be augmented:

$$\frac{dq_i}{dt}=w_3 \quad \frac{dc_l}{dt}=w_4 \quad \frac{df_p}{dt}=w_5 \tag{1c,1d,1e}$$

where q_i is the inlet flow rate, c_l is a parameter that characterizes tank leaks and f_p denotes the additional friction caused by partial blockage in the outlet pipeline. In this example, it is assumed that the means of the process noises w_is ($i=3, 4, 5$) are zero and the corresponding covariance matrix Q used in the Kalman filter is of the form:

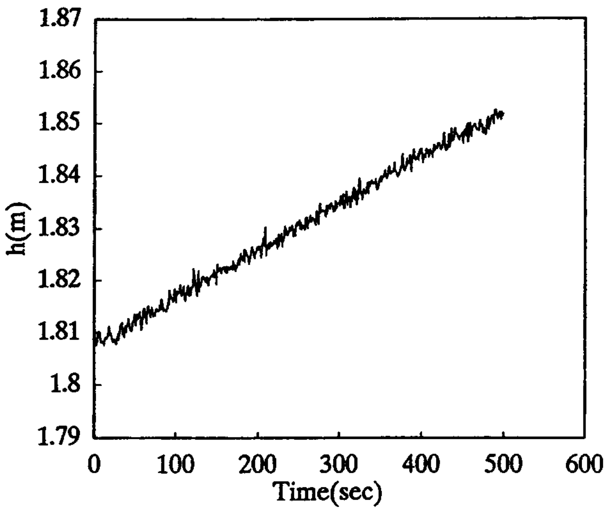

Figure 2a. Estimates generated by a conventional extended Kalman filter for the system in Example 1: the height of liquid level *h*.

$$Q=\begin{bmatrix} Q_{11} & 0 & 0 & 0 & 0 \\ 0 & Q_{22} & 0 & 0 & 0 \\ 0 & 0 & Q_{33} & 0 & 0 \\ 0 & 0 & 0 & Q_{44} & 0 \\ 0 & 0 & 0 & 0 & Q_{55} \end{bmatrix}$$

where the diagonal elements $Q_{ii}(i=1, \ldots, 5)$ are adjustable parameters.

In this example, it is also assumed that the height of the liquid level and the outlet flow rate can be measured directly and the means of the corresponding measurement noises are both zero. The variances of the measurement noises of h and q are assumed to be 10^{-6} and 10^{-10} respectively. The covariance matrix $\boldsymbol{R}$ adopted in the Kalman filter is also

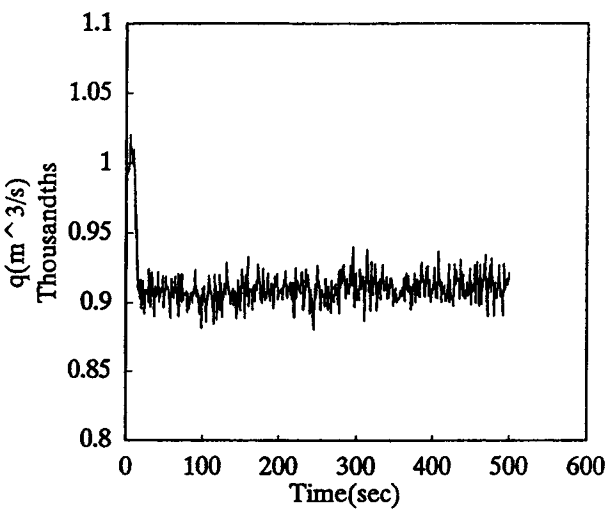

Figure 2b. Estimates generated by a conventional extended Kalman filter for the system in Example 1: the volumetric flow rate in the outlet pipeline *q*.

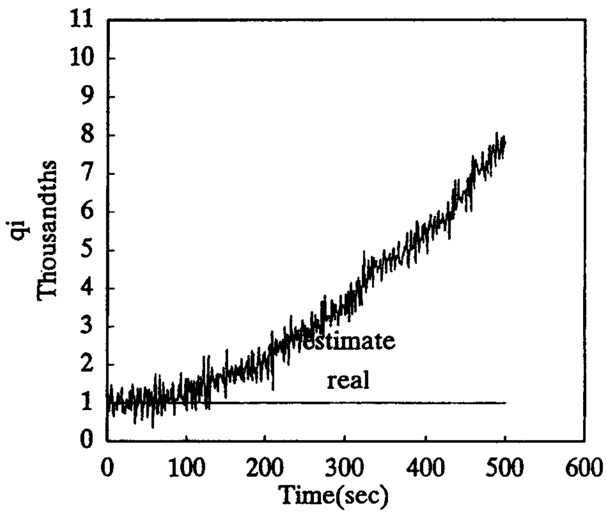

Figure 2c. Estimates generated by a conventional extended Kalman filter for the system in Example 1: the volumetric flow rate in the inlet pipeline q_i.

$$\boldsymbol{R}=\begin{bmatrix} 10^{-6} & 0 \\ 0 & 10^{-10} \end{bmatrix}$$

To test the effectiveness of the EKF, transient values of the state variables were generated by integrating Eqs. 1a and 1b with an equation that describes the dynamic behavior of a fictitious fault, that is, the outlet pipeline is plugged 10 s after the start of operation. This equation can be written as

$$\frac{df_p}{dt}=0.002[u(t-10)-u(t-15)] \tag{2}$$

where $u(t)$ is a unit step function. Before the inception of this

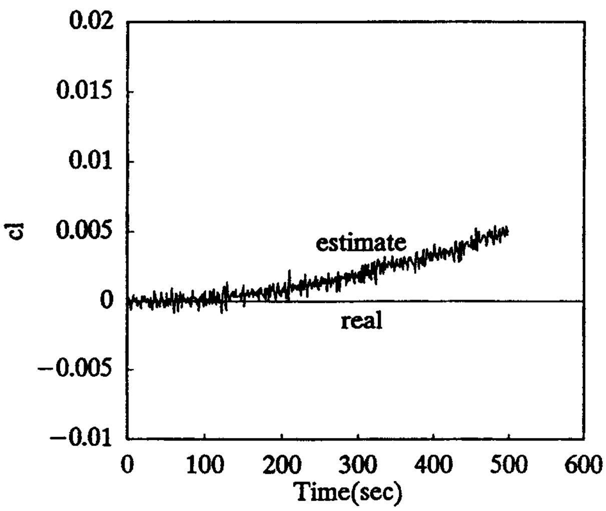

Figure 2d. Estimates generated by a conventional extended Kalman filter for the system in Example 1: the parameter associated with tank leaks c_l.

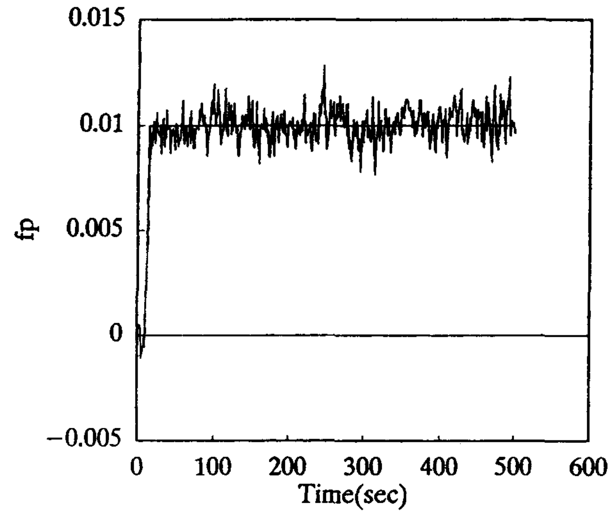

Figure 2e. Estimates generated by a conventional extended Kalman filter for the system in Example 1: the friction factor associated with partial blockage in the outlet pipeline f_p.

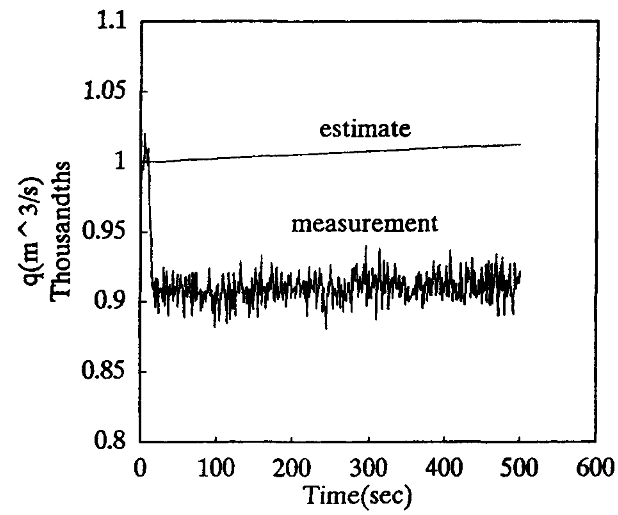

Figure 3b. Estimates generated by implementing EKFs in parallel for the system in Example 1: the volumetric flow rate in the outlet pipeline q.

fault, the system is at steady state. The steady-state values of the five state variables in Eqs. 1a–1e can be determined:

$$h = 1.808 \quad q = 0.001 \quad q_i = 0.001 \quad c_l = 0.0 \quad f_p = 0.0$$

They were used as the initial values in the simulation studies. After completing the numerical integration process, white Gaussian noises with variances 10^{-6} and 10^{-10} were then added to h and g to get the simulated measurement data.

Assuming the on-line measurements are taken once every second, one can obtain the estimates of all five state variables in Eqs. 1a–1e by a conventional EKF. Notice that the values of the diagonal elements in the covariance matrices Q need to be adjusted to optimize the performance of the EKF. The results presented in Figures 2a–2e were obtained after considerable effort in "tuning" the Kalman filter. From these results, one can observe that, although the estimates of h, q and f_p follow the simulated "real" values closely, those corresponding to q_i and c_l are still biased. This is certainly unacceptable for the purpose of fault diagnosis.

On the other hand, if the method of parallel parameter estimation is adopted, such drawbacks can be avoided completely. Using the same set of simulated measurement data, three smaller EKFs were implemented in parallel. In each EKF, only three state variables were included, that is, the height of liquid level h, the outlet flow rate q and one of the three parameters (q_i, c_l or f_p). Except for the one used as the state variable, the other two parameters were set to their initial values and assumed to be constants. If the anticipated fault is the actual one, the values of the parameters associated with other faults must stay unchanged by assumption and thus the esti-

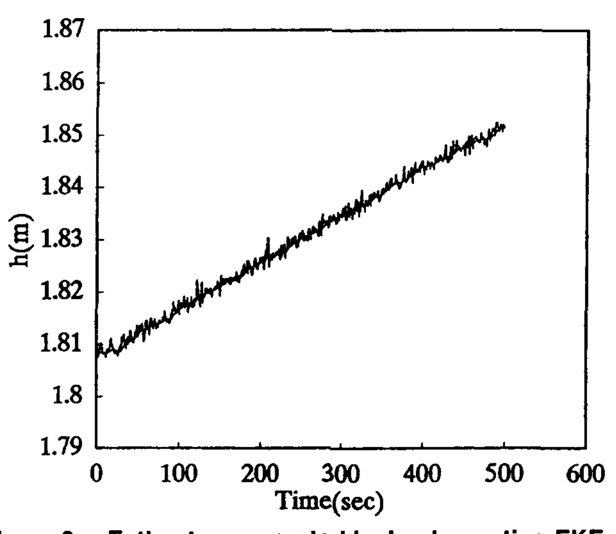

Figure 3a. Estimates generated by implementing EKFs in parallel for the system in Example 1: the height of liquid level in the tank h.

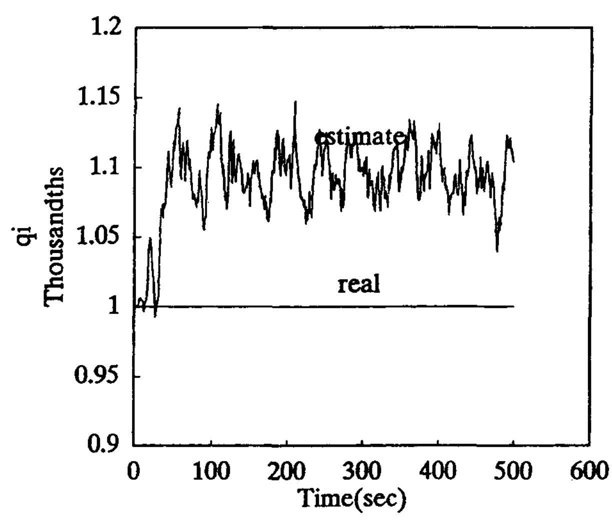

Figure 3c. Estimates generated by implementing EKFs in parallel for the system in Example 1: the volumetric flow rate in the inlet pipeline q_i.

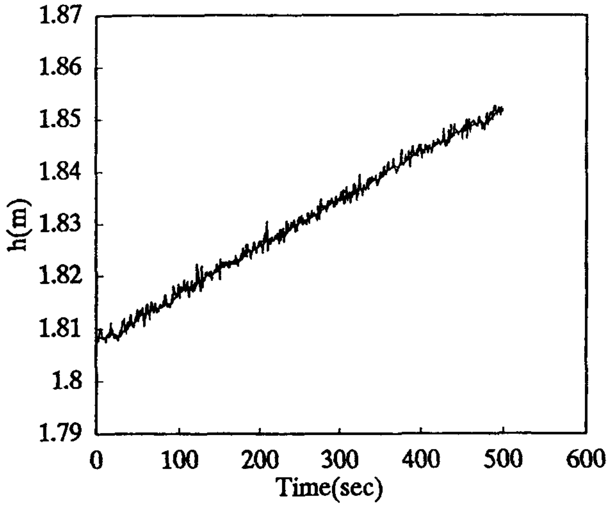

Figure 4a. Estimates generated by implementing EKFs in parallel for the system in Example 1: the height of liquid level in the tank *h*.

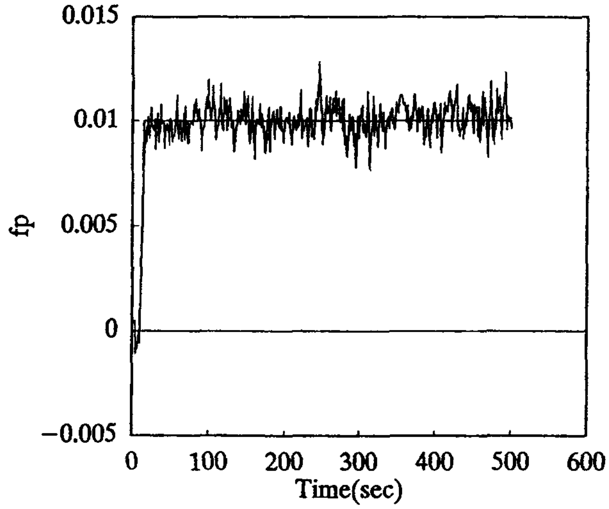

Figure 4c. Estimates generated by implementing EKFs in parallel for the system in Example 1: the friction factor associated with partial blockage in the outlet pipeline f_p.

mates of the state and augmented variables should be close to their correct values. Also, since the conclusions of diagnosis are, in general, relatively insensitive to changes in the magnitude of the variances used in the EKF, there is almost no need for tuning.

If one assumes that a change in inlet flow rate is the fault that causes abnormal system behavior, then the results presented in Figures 3a–3c can be obtained by the corresponding EKF calculations. Based on the fact that the estimates of q differ significantly from the measurement data (Figure 3b), one can then exclude the anticipated fault, that is, variation in q_i, from the list of candidates. Along the same reasoning, one can also dismiss the possibility of malfunctions associated with parameter c_l. However, if f_p is selected as the augmented state variable in EKF, then the estimates of h and q both trace the measurement data closely (Figure 4a and Figure 4b). According to these observations, one can correctly reach the conclusion that a partial blockage in the outlet pipeline is the origin of the fault. Furthermore, from the results presented in Figure 4c, one can clearly see that a good approximation of f_p was also obtained.

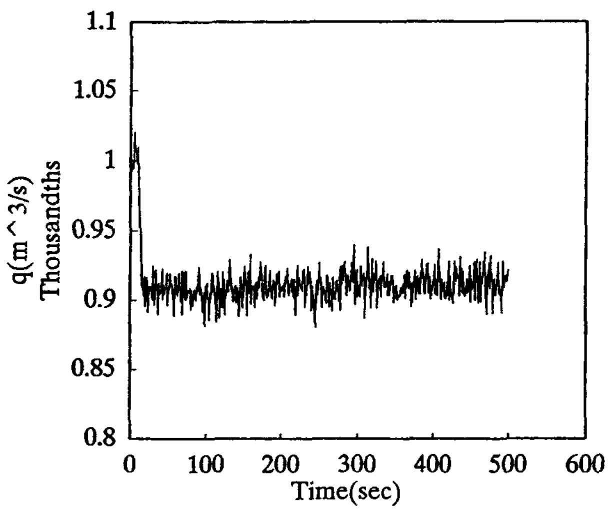

Figure 4b. Estimates generated by implementing EKFs in parallel for the system in Example 1: the volumetric flow rate in the outlet pipeline *q*.

Diagnostic Performance Table

If the parallel parameter estimation method is adopted in a fault monitoring system, its performance is dependent upon the choice of measurement variables. Even under the condition that only one fault occurs, inconsistent conclusions may still be drawn from different sets of on-line measurement data. Generally speaking, numerical simulation is one of the most commonly-used tools to assess the effects of varying the distribution of sensor locations. In this research, simulation studies have been carried out for all possible cases in every example. The corresponding results were summarized in a *diagnostic performance table*, which can be used as a basis for evaluating alternative monitoring systems. Three simple examples are presented in this section for the purpose of illustrating the compilation of such a table and to demonstrate some of the interesting phenomena that can be observed therein.

Example 2. Let us again consider the system described in Example 1. Assume that three parameters in the system model, q_i, c_l, and f_r, are subject to change during operation. Here f_r is a parameter which characterizes the effects of additional friction caused by a change in the valve position and/or partial blockage in the pipeline, that is, $f_r = f_v + f_p$. Simulation studies were first carried out under the condition that only one of the two state variables (h or q) could be measured on-line. The corresponding results are presented in two diagnostic performance tables, Table 1a and Table 1b. Notice that the chosen measurement variables are indicated at the upper-left corner of the diagnostic performance table. For each possible fault, simulated data were generated by introducing a change in the corresponding parameter. The entries in each row of the table were obtained from the same set of simulated measurement data corresponding to the parameter indicated at the left side

Table 1a. Diagnostic Performance of a One-Tank System with One Measured Variable

h	q_i	c_l	f_r
q_i	√	□	o
c_l	□	√	o
f_r	o	o	√

of the row. Each entry in this table is a conclusion drawn from analyzing the results of implementing one EKF to the above data. The augmented state variable adopted in this EKF is indicated on the top of the corresponding column. The symbols used in the diagnostic performance table are defined as follows:

• Diagonal Entries: The malfunction is correctly anticipated, that is, the augmented state variable is the parameter that actually deviates from the normal level. There are two possibilities:

√: The estimates of the parameter and the state variables match their actual transient behaviors.

*: The EKF fails to trace the variation in the assumed parameter.

• Off-Diagonal Entries: The anticipated fault does not occur, that is, the augmented state variable actually remains at the normal level during operation. Three types of observations may be obtained from the simulation results:

□: The estimates of all state variables (including measured and unmeasured variables) are good approximations of the simulated data. Thus, the corresponding fault cannot be excluded from the list of candidates.

o: Although the estimates of the unmeasured state variables may be significantly different from their actual behaviors, those of the measured variables follow the measurement values closely. Therefore, such a fault cannot be excluded.

blank: The estimates of some of the measured variables do not match the measurement values or the estimate of parameter remains unchanged. In this case, the corresponding fault can be regarded as nonexistent.

Obviously, the ideal sensor distribution is the one associated with a diagnostic performance table that contains only "√" (on the diagonal positions) and blanks (on the off-diagonal positions). In other words, to optimize the performance of such systems, the number of "*," "□" and "o" must be minimized. To avoid the errors "*," one must ensure that the effects of all possible faults can be observed on-line. In this study, this criterion is referred to as the *diagnostic observability* condition. A detailed discussion of this condition will be presented in the next section. The error "□" is caused by the fact that the effects of the actual fault on all the state variables can be reproduced exactly by the anticipated fault.

Table 1b. Diagnostic Performance of a One-Tank System with One Measured Variable

q	q_i	c_l	f_r
q_i	√	□	o
c_l	□	√	o
f_r	o	o	√

Table 1c. Diagnostic Performance of a One-Tank System with Two Measured Variables

h,q	q_i	c_l	f_r
q_i	√	□	
c_l	□	√	
f_r			√

Thus, from a structural standpoint, the parameters corresponding to these two faults are replaceable in the system model. For example, the parameters q_i and c_l appear only in Eq. 1a and, thus, both of them can only directly affect one state variable h which, in turn, can affect the other state variable q in Eq. 1b. As a result, the abnormal transient behaviors of either h or q, or both, which are caused by a decrease in the inlet flow rate q_i may be mistakenly regarded as due to tank leaks and, similarly, the opposite situation may also occur. This type of mistake clearly cannot be eliminated by measuring more state variables. The only way to solve this problem is to measure one (or more) of these parameters directly if possible.

The error "o" is due to the fact that the effects of the actual fault on the measured state variables can be replaced by those of the assumed fault. These errors can be removed by redistributing the sensor locations or placing additional measurement points. For example, if both state variables in the single-tank system are measured, all mistakes associated with "o" can be avoided in the corresponding diagnostic performance table (Table 1c).

In this research, the diagnostic observability criterion is considered to be a basic condition that all fault monitoring systems have to satisfy. Thus, diagnostic performance is essentially dependent only upon resolution. As mentioned before, the term diagnostic resolution in this article refers to a degree of uniqueness achieved in diagnosis. In fact, this quantity can be considered to be proportional to the number of blanks in the diagnostic performance table. Since the feasibility of measuring the parameters that produce "□" cannot always be guaranteed, the emphasis of this work is to develop a strategy for selecting measurement variables so that the number of "o" can be minimized in the resulting table. To demonstrate the effects of varying sensor distributions on resolution, two additional examples are provided:

Example 3. Let us consider the system shown in Figure 5, which consists of two tanks in series. The mathematical formulations of the models corresponding to these two tanks are essentially the same as Eq. 1a. Similarly, flow rates in the exit pipelines of these two tanks can be described, in principle, by Eq. 1b. However, in the model of pipeline 1, the liquid level h_1 should be replaced by the difference between the liquid levels in these two tanks, that is, $h_1 - h_2$.

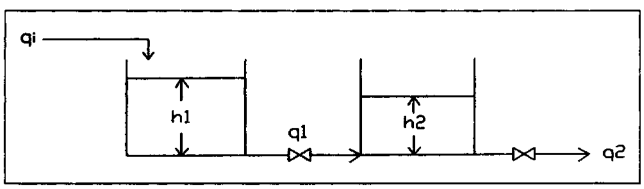

Figure 5. The simplified process flow diagram of a two-tank system.

Table 2a. Diagnostic Performance of a Two-Tank System (Figure 5) with Two Measured Variables

h_1,q_1	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√	o	o
c_{l2}			o	√	o
f_{r2}			o	o	√

Table 2b. Diagnostic Performance of a Two-Tank System (Figure 5) with Two Measured Variables

h_2,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□	o	o	
c_{l1}	□	√	o	o	
f_{r1}	o	o	√	o	
c_{l2}	o	o	o	√	
f_{r2}					√

Table 2c. Diagnostic Performance of a Two-Tank System (Figure 5) with Two Measured Variables

h_1,h_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√		
c_{l2}				√	o
f_{r2}				o	√

Table 2d. Diagnostic Performance of a Two-Tank System (Figure 5) with Two Measured Variables

h_1,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√		
c_{l2}				√	
f_{r2}					√

Table 2e. Diagnostic Performance of a Two-Tank System (Figure 5) with Two Measured Variables

q_1,h_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□	o		
c_{l1}	□	√	o		
f_{r1}	o	o	√		
c_{l2}				√	o
f_{r2}				o	√

Table 2f. Diagnostic Performance of a Two-Tank System (Figure 5) with Two Measured Variables

q_1,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□	o		
c_{l1}	□	√	o		
f_{r1}	o	o	√		
c_{l2}				√	
f_{r2}					√

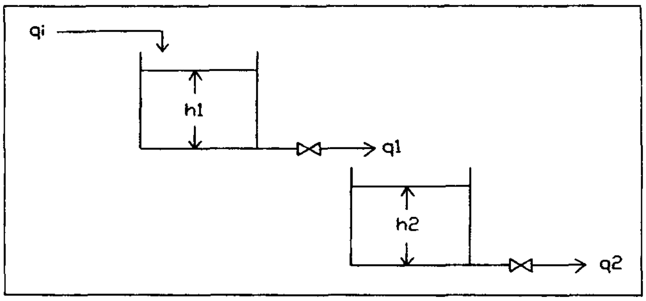

Figure 6. The simplified process flow diagram of a two-tank system.

There are four state variables, h_1, q_1, h_2, and q_2, in this problem. If only two state variables are allowed to be measured, then six combinations are possible. In this example, simulation studies have been carried out for all of them. The corresponding results are presented in Tables 2a-2f. As indicated before, since the error "□" can only be eliminated by measuring the parameters directly (which may not be feasible), the performance of the fault monitoring systems can be easily evaluated by counting the number of " o " in the above tables. Thus, one can see clearly that the best choice for this example is (h_1, q_2).

Example 4. Let us next consider a system with two tanks connected in a slightly different configuration (Figure 6). The mathematic formulations of both tanks and both pipelines are essentially the same as Eq. 1a and Eq. 1b respectively. If one considers the cases of measuring two state variables, then again six combinations need to be studied. However, the selection (h_1, q_1) should be excluded in this analysis due to its obvious violation of the diagnostic observability criterion. The results corresponding to the rest of the combinations are listed in Tables 3a-3e. From these tables, one can conclude that the best choices are (h_1, h_2) and (h_1, q_2). However, both of them still contain two errors associated with " o ." To remove these errors completely, additional measurement variables must be introduced into the sets of two state variables. Four possibilities need to be considered: (h_1, h_2, q_2), (q_1, h_2, q_2), (h_1, q_1, h_2) and (h_1, q_1, q_2). By means of simulation studies, the corresponding diagnostic performance tables can be produced (Tables 4a-4d). From these results, one can see that the optimal choice is (h_1, h_2, q_2).

In summary, one can conclude that system performance can be improved by varying the number and/or distribution of measurement points and, also, can be optimized by measuring only a portion of the entire set of state variables. On the other hand, it is quite obvious that, although the diagnostic performance tables are indeed helpful for the design of fault monitoring systems, the numerical effort needed to produce these tables can be overwhelming. Thus, there is a real incentive for developing a set of simple and effective design guidelines to replace the tedious simulation studies described in this section.

Diagnostic Observability

Let us now consider the standard form of a lumped model, which can be formulated as a set of nonlinear ordinary differential equations:

Table 3a. Diagnostic Performance of a Two-Tank System (Figure 6) with Two Measured Variables

h_2,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□	○	○	
c_{l1}	□	√	○	○	
f_{r1}	○	○	√	○	
c_{l2}	○	○	○	√	
f_{r2}					√

Table 3b. Diagnostic Performance of a Two-Tank System (Figure 6) with Two Measured Variables

h_1,h_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√		
c_{l2}				√	○
f_{r2}				○	√

Table 3c. Diagnostic Performance of a Two-Tank System (Figure 6) with Two Measured Variables

h_1,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√		
c_{l2}				√	○
f_{r2}				○	√

Table 3d. Diagnostic Performance of a Two-Tank System (Figure 6) with Two Measured Variables

q_1,h_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□	○		
c_{l1}	□	√	○		
f_{r1}	○	○	√		
c_{l2}				√	○
f_{r2}				○	√

Table 3e. Diagnostic Performance of a Two-Tank System (Figure 6) with Two Measured Variables

q_1,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□	○		
c_{l1}	□	√	○		
f_{r1}	○	○	√		
c_{l2}				√	○
f_{r2}				○	√

$$\frac{dx_1}{dt}=f_1(x_1, x_2, \cdots, x_n; c_1, c_2, \cdots, c_m)$$

$$\frac{dx_2}{dt}=f_2(x_1, x_2, \cdots, x_n; c_1, c_2, \cdots, c_m)$$

$$\vdots$$

$$\frac{dx_n}{dt}=f_n(x_1, x_2, \cdots, x_n; c_1, c_2, \cdots, c_m) \tag{3}$$

Table 4a. Diagnostic Performance of a Two-Tank System (Figure 6) with Three Measured Variables

q_1,h_2,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□	○		
c_{l1}	□	√	○		
f_{r1}	○	○	√		
c_{l2}				√	
f_{r2}					√

Table 4b. Diagnostic Performance of a Two-Tank System (Figure 6) with Three Measured Variables

h_1,q_1,h_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√		
c_{l2}				√	○
f_{r2}				○	√

Table 4c. Diagnostic Performance of a Two-Tank System (Figure 6) with Three Measured Variables

h_1,q_1,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√		
c_{l2}				√	○
f_{r2}				○	√

Table 4d. Diagnostic Performance of a Two-Tank System (Figure 6) with Three Measured Variables

h_1,h_2,q_2	q_i	c_{l1}	f_{r1}	c_{l2}	f_{r2}
q_i	√	□			
c_{l1}	□	√			
f_{r1}			√		
c_{l2}				√	
f_{r2}					√

where x_is ($i=1, 2, ..., n$) denote the state variables, c_js ($j=1, 2, ..., m$) represent the model parameters and f_ks ($k=1, 2, ..., n$) are some nonlinear continuous functions of x_is and c_js. Assume that the system is initially at steady state under normal operating conditions. Abnormal transient behaviors are caused only by changes in the parameters c_js. These changes can be viewed as the inputs to the system and the outputs are the state variables measured on-line. If any of the parameters deviates from its normal value, the resulting effects must show up first in the state variable(s) corresponding to the equation(s) in which this parameter appears. These effects then propagate to other parts of the system, that is, several other state variables are also affected later. Thus, to estimate the parameter associated with the assumed fault, one simply has to measure one of the state variables which are affected directly or indirectly by this parameter. Due to the single-fault assumption made in this study, good estimates can be obtained if the assumed fault is the correct one. Therefore, an *observable system* is defined as the one in which the effects of every possible fault can be detected from the chosen set of measurement variables.

Naturally, the question to ask is: how to choose these variables? To answer this question, one first needs to establish the precedence order of influences among state variables. For example, the variable h_2 in Example 4 can be affected by a change in the variable h_1, but the variation in h_2 may not create any effect on h_1. In this case, the change in h_1 (if it occurs) should always precede that of h_2. In simple systems, this precedence order may be determined directly from the corresponding process flow diagram. However, when complex units (such as a reactor) and/or complex configurations (such as multiple recycle loops) are involved, this approach is no longer feasible. Further, the idea of precedence order is also useful in establishing design guidelines for enhancing the diagnostic resolution of fault monitoring systems. Thus, there are enough incentives for developing a systematic procedure to establish this order for any model that can be written in the form of Eq. 3.

In this study, a partitioning algorithm (Steward, 1965) used for solving nonlinear algebraic equations was modified to achieve this purpose. Generally speaking, the best approach to promoting the computation efficiency in solving a large set of complex algebraic equations is to reorder them according to the structure of the system. The result of the reordering is called a *partition*. More specifically, a partition is the division of the set of equations into subsets, which we call *blocks*, so that each block in the partition is the smallest set of equations that must be solved simultaneously. After such a partition is established, the blocks can then be solved one at a time in series. In this research, the variable associated with time derivative in each of the model equations is treated as the output of the equation. This output variable is, of course, affected by the other variables which are included in the same equation. Thus, the functional relationships in Eq. 3 can also be described qualitatively in a *structural matrix* similar to the one used in the partitioning algorithm (Steward, 1965). The corresponding partition in this situation reflects the precedence order of influences among state variables.

Details of the modified partitioning algorithm are presented in the Appendix. To illustrate its implementation procedure, two examples are provided here:

Example 5. Let us consider the two-tank system described in Example 4. The corresponding structural matrix is shown in Figure 7a. Since this matrix is already in a *block triangular form*, this system can be directly divided into two blocks. The first block contains h_1 and q_1, while the other variables (h_2 and q_2) are included in the second block. Based on the specific form of this structural matrix, a *precedence diagram* can be constructed accordingly (Figure 7b). This diagram provides a clear picture of the precedence order of the changes that may occur in the state variables, that is, a change in one of the two variables in the first block generates deviations in both of the variables in the second block, but the variables in the first block cannot be affected by a fault that causes variations in the second block. In other words, the effects of a fault in tank 1 can be detected by measuring any one of the four state variables. However, the faults that occur in tank 2 are unobservable if one or both of the variables of tank 1 (h_1 and q_1) are selected for measurement. Thus, to ensure the observability

Variable

Equation	h_1	q_1	h_2	q_2
$\dot{h}_1$	×	×		
$\dot{q}_1$	×	×		
$\dot{h}_2$		×	×	×
$\dot{q}_2$			×	×

Figure 7a. The structural matrix corresponding to the two-tank system shown in Figure 6.

of the system, at least one of the variables in the second block must be included in the fault monitoring system.

Example 6. A more complex system is considered in this example (Figure 8a). There are 12 state variables (x_i, $i=1, 2, \ldots, 12$) here. x_1 to x_5 represent the heights of the liquid levels in 5 different tanks and x_6 to x_{12} denote the volumetric flow rates in their respective pipelines. By implementing the modified partitioning algorithm described in the Appendix, the structural matrix can be reordered in the block triangular form (Figure 8b). The corresponding precedence diagram is presented in Figure 8c. Notice that there are two branches in this diagram. From the most upstream variables, x_1, x_6 and x_7, the effects of a fault can propagate to x_2 and x_8 via the first branch and, through the second branch, to x_3, x_4, x_9, x_{10} and x_{11} and then to x_5 and x_{12}. This precedence order implies that at least two measurement variables, one in the end block of each branch, are needed in the fault monitoring system to satisfy the diagnostic observability condition.

One conclusion can be drawn from the above discussions: the diagnostic observability of a system is ensured by selecting one measurement variable in the end block of each branch of the precedence diagram. If a system consists of only one block, then this block itself is the "end" block. With such a selection strategy, unbiased estimates can be obtained under the condition that the assumed fault is the one that actually occurs, that is, all diagonal entries in the diagnostic performance table should be "√." However, identifying the correct fault origin is exactly the purpose of diagnosis and thus all possible faults must be considered in parallel. As a result, the diagnostic resolution may be unacceptably low, that is, the number of "o" is large, due to the fact that more than one fault may create the same abnormal behaviors observed in the measure-

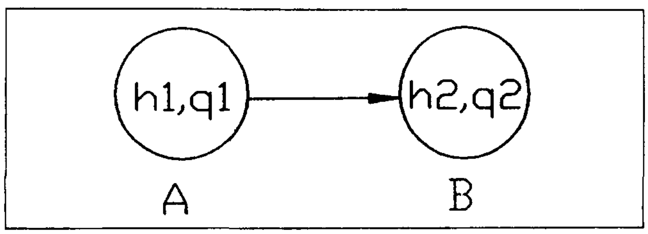

Figure 7b. The precedence diagram corresponding to the structural matrix shown in Figure 7a.

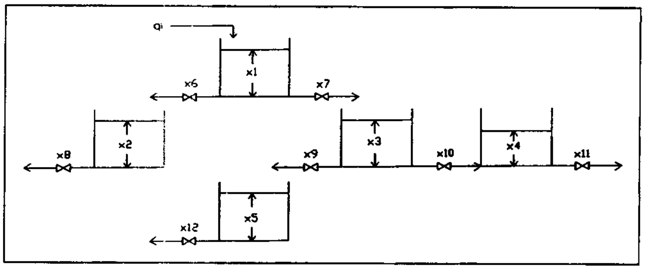

Figure 8a. The simplified process flow diagram of a 5-tank system.

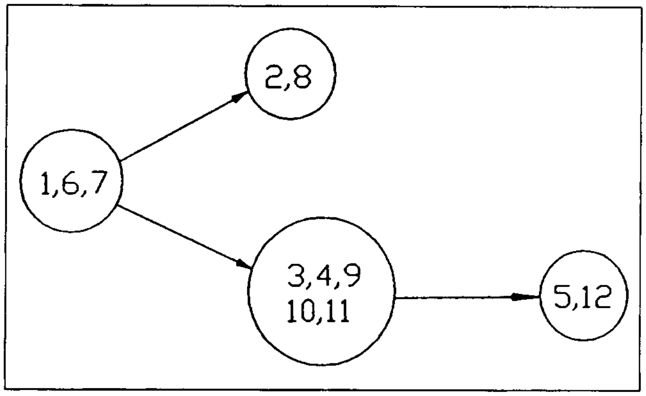

Figure 8c. The precedence diagram corresponding to the structural matrix shown in Figure 8b.

ment data. This problem will be analyzed in detail and the corresponding remedial measures will be developed in the next section.

Diagnostic Resolution

Reasoning from the fact that the number of " o " can be reduced by varying the distribution of sensor locations, one can clearly see that diagnostic resolution is related to the structure of the system model. The impact of noise is actually relatively unimportant. It can be shown that, even if the corresponding variances used in the EKFs are changed considerably, the resulting diagnostic performance table remains the same. Thus, the effects of the noise were neglected and only those concerning structural properties of models were analyzed in this article.

As a result of the optimization formulation used in deriving the Kalman filters, the estimated parameter obtained by the EKF algorithm usually forces the estimated values of the state variables to follow the measurement data as closely as possible. In an observable system, if the assumed fault in EKF is correctly anticipated, the results produced by the corresponding calculations can be verified by:

- treating the estimates of the parameter and the measurement variables as inputs to the model equations assumed in the EKF and, then,
- integrating the assumed model equations to determine the transient behaviors of all state variables.

The "estimates" of both the measured and unmeasured state variables generated in this fashion will, of course, be in close agreement with their actual values.

If the assumed fault in an EKF is different from the actual one, then several possible situations may occur after implementing the same procedure mentioned above. Firstly, the resulting estimates of all the state variables may still be good approximations of the true values. This fault then cannot be excluded. Thus, the corresponding entry in the diagnostic performance table should be filled with an " □ ." Secondly, the estimates of the measured variables may match the measurement data (their true values) well, but those of the unmeasured variables deviate from their actual behaviors. In this case, the assumed fault cannot be excluded either and should be classified as an error denoted by " o ." Finally, the calculated values of the measured variables may be significantly different from on-line data. On the basis of this observation, one can delete the assumed fault from the list of candidates. This is the situation corresponding to a blank in the diagnostic performance table.

From the above analysis, it can be seen that the diagnostic performance table can also be produced by the alternative procedure mentioned above. Notice that, if this alternative procedure is carried out without modifications, one still needs to carry out separate simulation studies, one for each entry, to construct the entire diagnostic performance table. However, these computations can be completely avoided if one takes advantage of the fact that, corresponding to the same set of chosen measurement variables, the calculation steps in the alternative procedure are the same for all entries in the diagnostic performance table. Thus, the key to this problem is analyzing the precedence order of the computation process described above. Under the condition that the system is observable (which can be created by the method outlined in the previous section), the off-diagonal entries in a diagnostic performance table can be determined, without numerical computation, by the following simple procedure:

1. Construct the structural matrix corresponding to the system model.

Variable (columns); Equation (rows)

Equation	x_1	x_6	x_7	x_2	x_8	x_3	x_4	x_9	x_{10}	x_{11}	x_5	x_{12}
$\dot{x}_1$	×	×	×									
$\dot{x}_6$	×	×										
$\dot{x}_7$	×		×									
$\dot{x}_2$		×		×	×							
$\dot{x}_8$				×	×							
$\dot{x}_3$			×			×		×	×			
$\dot{x}_4$							×		×	×		
$\dot{x}_9$						×		×				
$\dot{x}_{10}$						×	×		×			
$\dot{x}_{11}$							×			×		
$\dot{x}_5$								×			×	×
$\dot{x}_{12}$											×	×

Figure 8b. The structural matrix corresponding to the 5-tank system shown in Figure 8a.

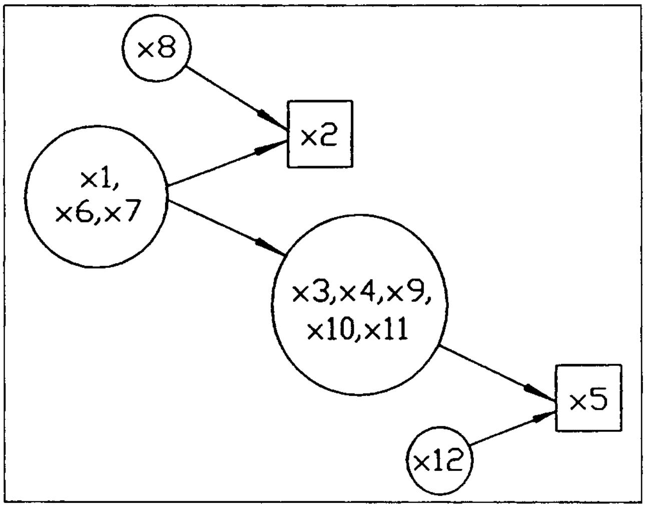

Figure 9a. The precedence order of the proposed verification procedure corresponding to two measured variables (x_2 and x_5) for the system shown in Figure 8a.

2. Eliminate the off-diagonal entries corresponding to the measurement variables in the structural matrix.

3. Apply the partitioning algorithm described in the Appendix to the matrix obtained in the previous step. Draw the resulting precedence diagram.

4. Select an end block in the precedence diagram which contains a measurement variable. Trace a path backward from this block until a branching block is reached. Remove the branching block from the path. Repeat this path-tracing procedure until all such paths are exhausted.

5. Repeat Step 4 until all such end blocks are exhausted.

6. Select one of the paths obtained in the previous step. List all the model equations on this path and also all the parameters in these equations. If there are more than two parameters in the resulting list and these parameters do not appear elsewhere, then fill the corresponding positions in the diagnostic performance table with the symbols " o ." Repeat this process until all such paths are exhausted.

7. If two or more parameters appear only in one equation,

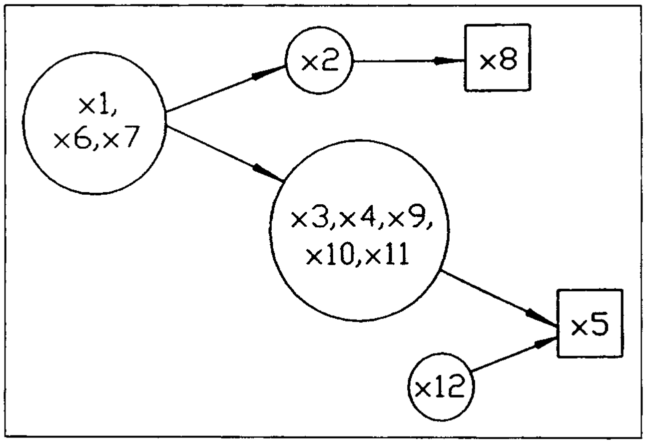

Figure 9c. The precedence order of the proposed verification procedure corresponding to two measured variables (x_5 and x_8) for the system shown in Figure 8a.

then place " □ " at the corresponding locations in the diagnostic performance table. Repeat this process until all such cases are exhausted.

8. The rest of the entries in the diagnostic performance table should be left blank.

Notice that, as a result of Step 2, unmeasured state variables are guaranteed not to appear in the terminal blocks of the paths obtained in Step 5. Each of these blocks contains one and only one measurement variable and no unmeasured variables. If two or more faults can affect the variables on the same path and the corresponding parameters do not appear elsewhere in the system model, then their effects can only be detected from the measurement variable in the end block of this path. Generally speaking, if one of these faults occurs during operation, the abnormal behavior observed in the measurement data can be interpreted as due to a change in any of the above parameters on the path. Thus, if the assumed fault is incorrect and, also, corresponding to one of these parameters, an error associated with " o " will be committed in diagnosis.

If the effects of a fault can be observed from two or more independent sources of measurement information, then in general this fault can be uniquely identified. In other words, the

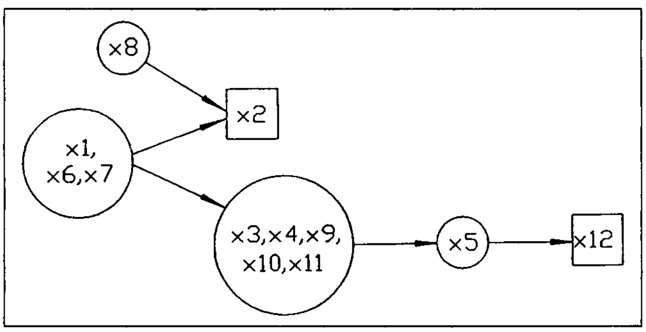

Figure 9b. The precedence order of the proposed verification procedure corresponding to two measured variables (x_2 and x_{12}) for the system shown in Figure 8a.

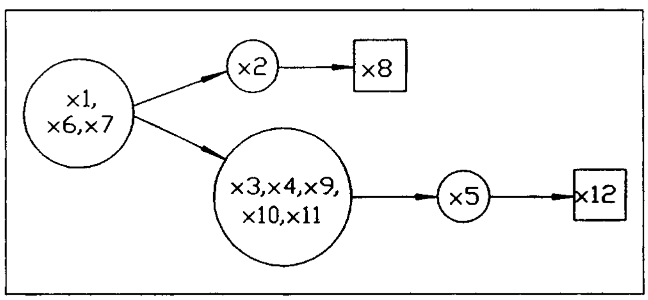

Figure 9d. The precedence order of the proposed verification procedure corresponding to two measured variables (x_5 and x_{12}) for the system shown in Figure 8a.

Table 5a. Diagnostic Performance of a 5-Tank System (Figure 8a) with Two Measured Variables

x_2, x_5	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√						o				
c_{l3}				√	o	o				o	o	o	
c_{l4}				o	√	o				o	o	o	
c_{l5}				o	o	√				o	o	o	o
f_{r1}							√						
f_{r2}								√					
f_{r3}			o						√				
f_{r4}				o	o	o				√	o	o	
f_{r5}				o	o	o				o	√	o	
f_{r6}				o	o	o				o	o	√	
f_{r7}						o							√

corresponding parameter is associated with a branching block, one of its upstream blocks or multiple blocks on more than one path. Also, although several different faults may affect the same set of measurement variables, it is assumed here that their effects on these variables are distinguishable in most realistic systems. Thus, one of the heuristic guidelines used in the above procedure for constructing the diagnostic performance tables is that the off-diagonal positions corresponding to these parameters should be left blank. There is, however, one exception to this rule: these parameters may appear only in one equation. In such a case, their effects on all measurement variables are interchangeable and the corresponding entries in the diagnostic performance table should be "□."

To illustrate the implementation procedure of the proposed method, the following example is provided:

Example 7. Let us continue our study on the system described in Example 6. From the conclusions of the previous example, two measurement variables, that is, one each in the blocks $\{x_2, x_8\}$ and $\{x_5, x_{12}\}$, are needed to ensure diagnostic observability. There are four combinations possible: (x_2, x_5), (x_2, x_{12}), (x_5, x_8) and (x_5, x_{12}). Their precedence diagrams can be obtained by carrying out Step 1 to Step 3 of the proposed procedure (Figure 9a-9d). From these diagrams, the corresponding diagnostic performance tables can be determined by implementing Step 4 to Step 8. We have found that the two diagnostic performance tables corresponding to (x_2, x_5) and (x_5, x_8) are exactly the same (Table 5a). Similarly, either one of the two alternatives, (x_2, x_{12}) and (x_8, x_{12}), can be selected to produce the results shown in Table 5b. From these two tables, one can conclude that the resolution of the former two cases is higher and thus they represent better choices if only two measurement variables are allowed.

Table 5b. Diagnostic Performance of a 5-Tank System (Figure 8a) with Two Measured Variables

x_2, x_{12}	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√						o				
c_{l3}				√	o	o				o	o	o	o
c_{l4}				o	√	o				o	o	o	o
c_{l5}				o	o	√				o	o	o	o
f_{r1}													
f_{r2}							√	√					
f_{r3}			o						√				
f_{r4}				o	o	o				√	o	o	o
f_{r5}				o	o	o				o	√	o	o
f_{r6}				o	o	o				o	o	√	o
f_{r7}				o	o	o				o	o	o	√

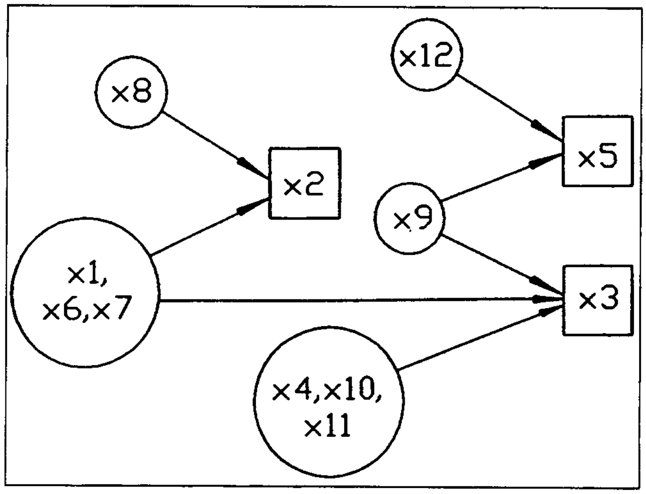

Figure 10a. The precedence order of the proposed verification procedure corresponding to three measured variables (x_2, x_3 and x_5) for the system shown in Figure 8a.

To improve the resolution of the fault monitoring system, additional measurement points must be introduced. Let us first consider the precedence diagram corresponding to the selection (x_2, x_5), that is, Figure 9a. One can clearly see that the major source of "o" comes from the block $\{x_3, x_4, x_9, x_{10}, x_{11}\}$. Thus, the natural choice will be one of these variables. The precedence diagrams and diagnostic performance tables corresponding to all five possible combinations are presented in Figures 10a-10e and in Tables 6a-6e, respectively. By counting

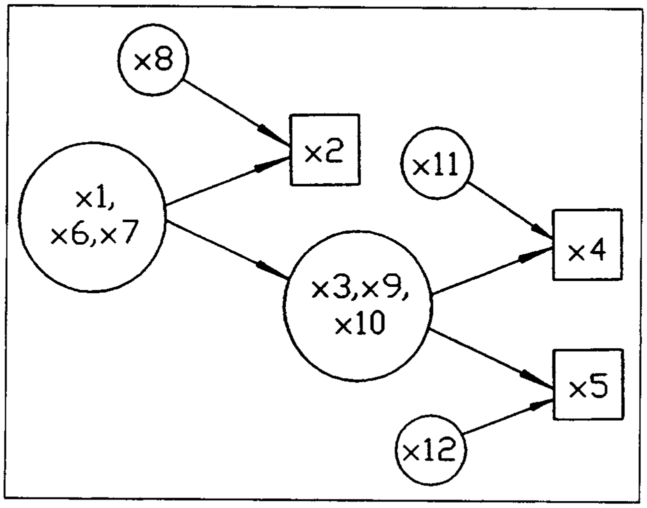

Figure 10b. The precedence order of the proposed verification procedure corresponding to three measured variables (x_2, x_4 and x_5) for the system shown in Figure 8a.

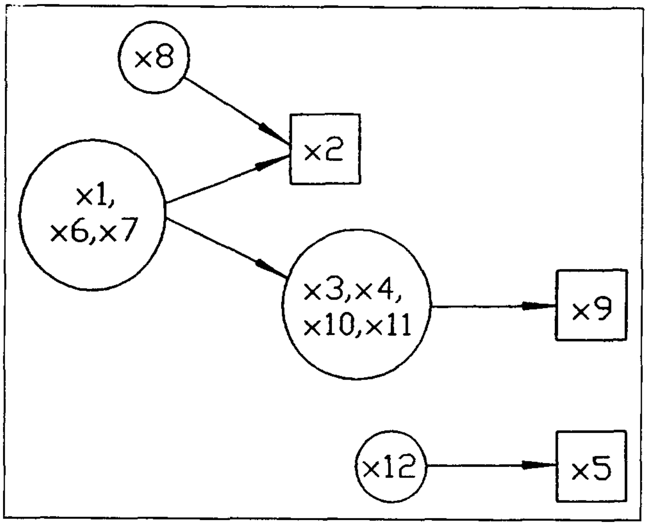

Figure 10c. The precedence order of the proposed verification procedure corresponding to three measured variables (x_2, x_5 and x_9) for the system shown in Figure 8a.

the number of " o " in these tables, one can see that the best choice is (x_2, x_5, x_{11}).

Based on Figure 10e, it is obvious that x_8 and x_{12} need to be measured to completely eliminate the errors associated with " o ." Thus, to minimize the number of " o "s in the diagnostic performance table for this example, five measurement variables are needed: x_2, x_5, x_8, x_{11} and x_{12}. Its precedence diagram and diagnostic performance table are given in Figure 11 and Table 7, respectively. Furthermore, if the measurement of the inlet flow rate to the tank located farthest upstream is feasible, the errors in diagnosis can be eliminated entirely. Finally, it should be pointed out that this set of measurement variables

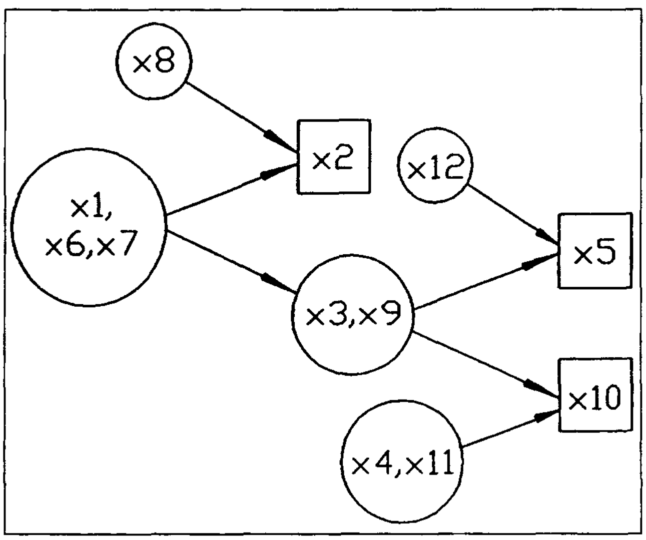

Figure 10d. The precedence order of the proposed verification procedure corresponding to three measured variables (x_2, x_5 and x_{10}) for the system shown in Figure 8a.

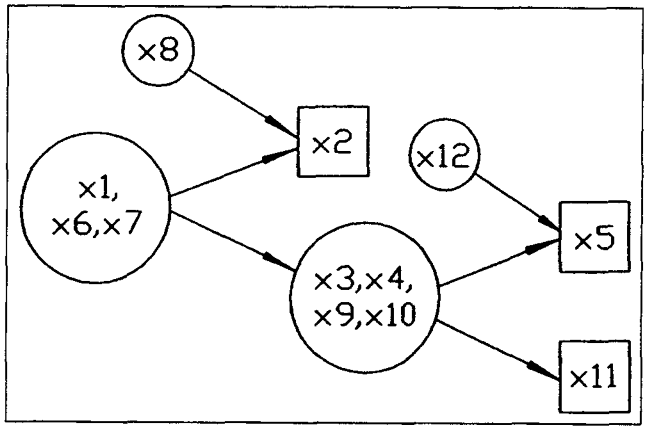

Figure 10e. The precedence order of the proposed verification procedure corresponding to three measured variables (x_2, x_5 and x_{11}) for the system shown in Figure 8a.

can also be obtained by considering other alternatives which also satisfy the diagnostic observability criterion, that is, (x_2, x_{12}), (x_5, x_8) and (x_8, x_{12}).

Notice that a diagnostic performance table can be constructed with the proposed algorithm for any given set of measurement variables. If the sensor for measuring a particular state variable is unavailable or too expensive, then this variable

Table 6a. Diagnostic Performance of a 5-Tank System (Figure 8a) with Three Measured Variables

x_2,x_3,x_5	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√						o				
c_{l3}				√	o						o	o	
c_{l4}				o	√						o	o	
c_{l5}						√							o
f_{r1}													
f_{r2}							√	√					
f_{r3}			o						√				
f_{r4}										√			
f_{r5}				o	o						√	o	
f_{r6}				o	o						o	√	
f_{r7}						o							√

Table 6b. Diagnostic Performance of a 5-Tank System (Figure 8a) with Three Measured Variables

x_2,x_4,x_5	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√						o				
c_{l3}				√									
c_{l4}					√							o	
c_{l5}						√							o
f_{r1}							√						
f_{r2}								√					
f_{r3}			o						√				
f_{r4}										√			
f_{r5}											√		
f_{r6}					o							√	
f_{r7}						o							√

Table 6c. Diagnostic Performance of a 5-Tank System (Figure 8a) with Three Measured Variables

x_2,x_5,x_9	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√						o				
c_{l3}				√	o					o	o	o	
c_{l4}				o	√					o	o	o	
c_{l5}						√							o
f_{r1}							√						
f_{r2}								√					
f_{r3}			o						√				
f_{r4}				o	o					√	o	o	
f_{r5}				o	o					o	√	o	
f_{r6}				o	o					o	o	√	
f_{r7}						o							√

Table 6d. Diagnostic Performance of a 5-Tank System (Figure 8a) with Three Measured Variables

x_2,x_5,x_{10}	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√						o				
c_{l3}				√									
c_{l4}					√						o	o	
c_{l5}						√							o
f_{r1}							√						
f_{r2}								√					
f_{r3}			o						√				
f_{r4}										√			
f_{r5}					o						√	o	
f_{r6}					o						o	√	
f_{r7}						o							√

Table 6e. Diagnostic Performance of a 5-Tank System (Figure 8a) with Three Measured Variables

x_2,x_5,x_{11}	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√						o				
c_{l3}				√									
c_{l4}					√								
c_{l5}						√							o
f_{r1}							√						
f_{r2}								√					
f_{r3}			o						√				
f_{r4}										√			
f_{r5}											√		
f_{r6}												√	
f_{r7}						o							√

can be excluded from consideration. The best fault monitoring system is simply the one with the maximum number of blanks in the diagnostic performance table.

It should be pointed out that the suggested procedure for selecting the measurement variables has been verified by numerous simulation studies. The diagnostic performance tables predicted by applying this procedure to the models described in Examples 2, 3 and 4 are exactly the same as those produced numerically in Tables 1a-1c, 2a-2f, 3a-3e and 4a-4d. Also, the correctness and effectiveness of the proposed approach has

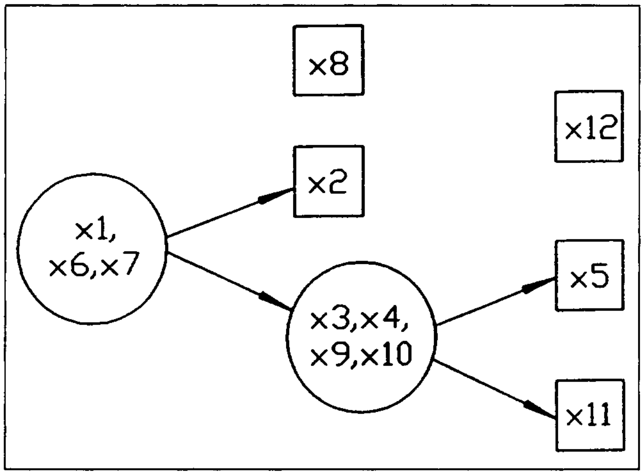

Figure 11. The precedence order of the proposed verification procedure corresponding to five measured variables (x_2, x_5, x_8, x_{11} and x_{12}) for the system shown in Figure 8a.

been demonstrated without exceptions in many other more complex examples, for example, a number of 3-tank systems in various different configurations (Mah, 1992) and the 5-tank process described in Example 7. In addition, since the development of the proposed approach is general, extensions to processes other than the tank systems is straightforward. A realistic example is provided as the supplementary material of this article. Thus, although the design strategy presented here is based only a qualitative analysis, due to its reliability and efficiency, this method should be considered as a practical tool for the placement of sensor locations in complex chemical plants.

Methods for Enhancing Diagnostic Performance

Although the number of " o " in the diagnostic performance table can be minimized by altering the distribution of sensor locations, mistakes associated with "□" or " o " are still possible in diagnosis. Thus, there is a need to develop additional performance-enhancing methods by making use of information which is not included in the previous analysis. Two types of such informations are useful, that is, the physical significances of the model parameters and the fault patterns obtained from the parallel estimation calculations.

Generally speaking, variations in the same parameter may be interpreted differently on the basis of their trends. For example, there are two possible failure modes associated with a valve, that is, the valve fails to open or close, and they can be described by changing the value of parameter f_v in opposite directions. On the other hand, there is no guarantee of one-to-one correspondence between the deviations in parameters and the actual faults. For example, the parameter associated with a tank leak, c_l, cannot possibly be negative. Thus, with the help of these insights, one should be able to eliminate some of the errors represented by " o " and "□" in the diagnostic performance table.

Although the number of blanks cannot be further increased after implementing the method suggested above, still higher diagnostic resolution can be realized via the analysis of fault

Table 7. Diagnostic Performance of a 5-Tank System (Figure 8a) with Five Measured Variables

$x_2,x_5,x_8,x_{11},x_{12}$	q_i	c_{l1}	c_{l2}	c_{l3}	c_{l4}	c_{l5}	f_{r1}	f_{r2}	f_{r3}	f_{r4}	f_{r5}	f_{r6}	f_{r7}
q_i	√	□											
c_{l1}	□	√											
c_{l2}			√										
c_{l3}				√									
c_{l4}					√								
c_{l5}						√							
f_{r1}							√						
f_{r2}								√					
f_{r3}									√				
f_{r4}										√			
f_{r5}											√		
f_{r6}												√	
f_{r7}													√

patterns. As mentioned before, the entries in each row of the diagnostic performance table are obtained by applying the parallel parameter estimation method to the measurement data collected on the same fault. It may be possible for several "○"s and/or "□"s to appear in the same row. However, their patterns corresponding to different faults may not be the same. Thus, by comparing the fault patterns in the diagnostic performance table, a fault may be uniquely identified even when the errors "○" and/or "□" still exist in the corresponding row.

The following example is presented to illustrate the usefulness of these two types of information:

Example 8. Let us make use of the results obtained in Example 3. Consider the case in which h_1 and q_1 are selected as the measurement variables, that is, Table 2a. Assume that there are 10 possible faults in this system, that is, a change in the inlet flow rate ($q_i\uparrow$ or $q_i\downarrow$), a leak in one of the tanks ($c_{l1}\uparrow$ or $c_{l2}\uparrow$), partial blockage in one of the pipelines ($f_{p1}\uparrow$ or $f_{p2}\uparrow$), accidental opening of one of the valves ($f_{v1}\downarrow$ or $f_{v2}\downarrow$) or accidental closure of one of the valves ($f_{v1}\uparrow$ or $f_{v2}\uparrow$). After taking into account of their effects on the parameters, Table 2a can be changed into Table 8a. In this table, a "×" is used to indicate the fact that the corresponding estimate does not have any physical meaning and can be removed from the table. Furthermore, since the fault origin is not known before diagnosis, the faults associated with "√," "□" or "○" are actually indistinguishable, that is, Table 8a should be replaced by Table 8b. The entries in each row of Table 8b represent the fault pattern corresponding to the fault origin. From row 2 and row 3 of this table, one can see that a decrease in the inlet flow rate may be confused with a leak in the first tank or vice versa. Similarly, from the observation that the fault patterns in several other groups of rows (4,5), (8,9) and (7,10) are the same, the corresponding faults in each group cannot be differentiated either. Thus, if any of these faults occurs, the results of diagnosis must include two candidates. Also, from rows 1 and 6, one can conclude that there are only two uniquely identifiable faults: those represented by $q_i\uparrow$ and $f_{v1}\downarrow$.

Finally, it should be pointed out that, even with these performance-enhancing methods, the best overall performance of the fault monitoring system can only be achieved by placing the measurement points according to the procedure suggested in the previous sections. For example, if the h_1 and q_2 are

Table 8a. Modified Diagnostic Performance of a Two-Tank System (Figure 5) with Two Measured Variables

h_1,q_1	q_i	c_{l1}	f_{p1}	f_{v1}	c_{l2}	f_{p2}	f_{v2}
$q_i\uparrow$	√	⊠					
$q_i\downarrow$	√	□					
$c_{l1}\uparrow$	□	√					
$f_{p1}\uparrow$			√	□	⊗	⊗	⊗
$f_{v1}\uparrow$			□	√	⊗	⊗	⊗
$f_{v1}\downarrow$			⊠	√	⊗	⊗	⊗
$c_{l2}\uparrow$			⊗	○	√	⊗	○
$f_{p2}\uparrow$			○	○	⊗	√	□
$f_{v2}\uparrow$			○	○	⊗	□	√
$f_{v2}\downarrow$			⊗	○	○	⊠	√

Table 8b. Fault Patterns for a Two-Tank System (Figure 5) with Two Measured Variables

h_1,q_1	q_i	c_{l1}	f_{p1}	f_{v1}	c_{l2}	f_{p2}	f_{v2}
$q_i\uparrow$	△						
$q_i\downarrow$	△	△					
$c_{l1}\uparrow$	△	△					
$f_{p1}\uparrow$			△	△			
$f_{v1}\uparrow$			△	△			
$f_{v1}\downarrow$				△			
$c_{l2}\uparrow$				△	△		△
$f_{p2}\uparrow$			△	△		△	△
$f_{v2}\uparrow$			△	△		△	△
$f_{v2}\downarrow$				△	△		△

measured in Example 8, then the faults corresponding to row 7 and row 10 of Table 8b can be differentiated without difficulties.

On-Line Implementation Strategies

The computational load of the parallel parameter estimation method is quite demanding if it is implemented literally without modifications. To satisfy the needs in practical applications, several time-saving strategies have been adopted in this work.

First of all, the differential equations adopted in each of the EKFs does not have to be associated with the model of the entire system. Only a subset of the model equations is really needed. This point can be seen by applying the partitioning procedure in the Appendix to decompose the system into several blocks. Since one specific fault is assumed to occur in each of the EKFs, it is only necessary to consider those blocks that are affected by the assumed fault, that is, the blocks in which the corresponding parameter appears and their downstream blocks. If some of the upstream variables are included in this subset of model equations, their values should be considered to be at the normal levels without variations. Notice that, in implementing these parallel EKFs, if a parameter is found to be significantly different from its normal value, then additional tests must be made to determine whether the assumed fault is indeed the correct one. More specifically, it is necessary to compare:

- the estimates and the measurement data of the downstream measured variables; and
- the assumed (normal) values and the measurement data of other measured variables.

A standard hypothesis testing technique (Himmelblau, 1978) can be utilized for this purpose. If any of the above tests fails, then the corresponding entry in the diagnostic performance table should be a blank. Otherwise, the assumed fault should be considered as one of the candidates.

From the above discussions, one can see that the sizes of the EKFs used in the parallel parameter estimation method do not have to be the same. Since the EKFs corresponding to the parameters in the downstream blocks are smaller, the effort in computation can be reduced considerably. But, it is still necessary to estimate all the parameters in the system model. To cut down the computational load further, a two-step technique can be utilized. First, the subsystem in which the actual fault occurs can be identified by comparing the measurement data with their normal values. Then the on-line estimation technique can be applied to a combined system which consists of this subsystem and all its downstream blocks. In this case, estimation of the parameters that appear outside the subsystem becomes unnecessary. Thus, one of the obvious advantages of the two-step approach is that the number of activated EKFs can be reduced significantly. Furthermore, the calculations involved in parameter estimation can be avoided completely during normal operation. Following is a simple procedure for identifying the subsystem in which the fault occurs:

1. Apply the partitioning algorithm described in the Appendix and draw the corresponding precedence diagram.
2. Find the blocks in which all the measured variables remain normal. Remove these blocks and their upstream blocks from the precedence diagram.
3. Identify the blocks farthest upstream in which one or more measured variables are located. These blocks and their upstream blocks form the subsystem in which the actual fault occurs.

Naturally, the above strategy is not without drawbacks. For example, the criteria for testing hypotheses have not been well established and thus there is a chance for misdiagnosis or false alarm. Also, not only the model equations associated with the subsystem but also those in its downstream blocks must be included in the EKF calculations. Therefore, more than one block may be included in the combined system (the subsystem and its downstream blocks) and the size of the corresponding model may still be very large. In implementing the parallel estimation method, each EKF is used to determine the value of the parameter corresponding to one possible fault in the subsystem only. Instead of applying the conventional algorithm directly, several smaller EKFs, each of them associated with one of the blocks, can be used in series to achieve the same purpose. EKFs are first applied to the blocks located farthest downstream. Upstream variables that appear in these blocks should be considered as the "parameter" to be estimated. The estimates of these variables can then be treated as the "measurement data" in their respective blocks. This process can be repeated until the block in which the assumed fault exists.

To illustrate this implementation procedure, a simple example is presented in below:

Example 9. Let us consider the system described in Example 4. This system can be separated into two blocks, that is, an upstream block $\{h_1, q_1\}$ and a downstream block $\{h_2, q_2\}$. If the anticipated malfunction is a leak in the upstream tank, then the model equations included in the EKF associated with the downstream block is

$$A_2\frac{dh_2}{dt}=q_1-q_2 \tag{3a}$$

$$\frac{dq_2}{dt}=\frac{\pi d_2^2}{4\rho l_2}\left[\rho g h_2-(f_2+f_{v2})\frac{8l_2\rho q_2|q_2|}{\pi^2 d_2^5}\right] \tag{3b}$$

$$\frac{dq_1}{dt}=w_1 \tag{3c}$$

Here, the outlet flow rate of the upstream tank q_1 is treated as a parameter to be estimated by the Kalman filter. The estimated values of q_1 can then be considered as the measurement data in the EKF associated with the upstream block. The corresponding model equations are:

$$A_1\frac{dh_1}{dt}=q_i-q_1-c_{l1}\sqrt{h_1} \tag{3d}$$

$$\frac{dq_1}{dt}=\frac{\pi d_1^2}{4\rho l_1}\left[\rho g h_1-(f_1+f_{v1})\frac{8l_1\rho q_1|q_1|}{\pi^2 d_1^5}\right] \tag{3e}$$

$$\frac{dc_{l1}}{dt}=w_2 \tag{3f}$$

This approach has been verified by simulation for various combinations of measurement variables. We have found that all conclusions in the diagnostic performance table remain unchanged.

For this problem, the savings in computation time is not obvious. Notice that the number of model equations included in a conventional EKF is five. Thus, the number of differential equations involved in actual computation should be 20. On the other hand, due to the fact that there are three model equations in each block and thus nine differential equations in each EKF, this number can be reduced to 18 if the proposed approach is adopted. However, if one considers a larger and more complex system, their difference will become more apparent. For example, let us consider a 10-tank system in which the tanks are connected in the same fashion as that of Figure 6. In this case, the number of differential equations involved in the conventional algorithm should be 252. But, only 90 equations are needed if the proposed method is adopted.

Conclusions

A simple yet practical design strategy has been developed in this study to optimize the performance of model-based fault monitoring systems. The success of such a strategy is built upon several new features which have not been implemented before. Firstly, a parallel parameter estimation method has been adopted to reduce the chance of producing biased estimates, which is a phenomenon commonly seen in the implementation of the extended Kalman filter. Secondly, to quantify the performance of fault monitoring systems, the concepts of *diagnostic observability* and *diagnostic resolution* have been introduced and represented by diagnostic performance tables. Thirdly, Steward's partitioning algorithm has been modified for the purpose of selecting the smallest set of measurement variables necessary for ensuring observability. Fourthly, to achieve the highest resolution, a simple procedure has been worked out for placing the sensor locations only on the basis of the structural information of the system model. Fifthly, other performance-enhancing measures have also been developed on the basis of additional insights gained by analyzing the physical significance of the model parameters and comparing the fault patterns in the diagnostic performance table. Finally, several on-line implementation strategies have been proposed to relieve part of the computation burden in carrying out the parallel EKFs.

It must be emphasized that, although the development of this design strategy is not theoretically rigorous, the correctness of its results has been supported by numerous simulation studies. In addition, since implementation is simple and easy, one can conclude that the proposed approach is practical even for large and complex chemical processes.

Acknowledgment

This work is supported by the national science council of the ROC under Grant NSC81-0414-P006-15B.

Notation

A = the cross-sectional area of a tank
c_l = the parameter that characterizes tank leaks
d = the diameter of a pipeline
f = the friction factor for the flow in a pipeline
f_p = the parameter that represents the additional friction caused by partial blockage in a pipeline
f_r = $f_p + f_v$
f_v = the parameter that represents the additional friction caused by the valve in a pipeline
h = the height of liquid level in a tank
l = the length of a pipeline
q = the volumetric flow rate in a pipeline
$\boldsymbol{Q}$ = covariance matrix corresponding to the system model
$\boldsymbol{R}$ = covariance matrix corresponding to the measurement noise
t = time

Greek letter

ρ = the density of liquid

Literature Cited

Chang, C. C., and C. C. Yu, "On-line Fault Diagnosis Using the Signed Directed Graph," *Ind. Eng. Chem. Res.*, **29**, 1290 (1990).

Churchill, S. W., "Friction-Factor Equation Spans all Fluid-Flow Regimes," *Chem. Eng.*, **84**(24), 91 (1977).

Dalle Molle, D. T., and D. M. Himmelblau, "Fault Detection in a Single-Stage Evaporator Via Parameter Estimation using the Kalman Filter," *Ind. Eng. Chem. Res.*, **26**, 2482 (1987).

Davis, J. F., S. K. Shum, B. Chandrasekaran, and W. F. Punch, III, "A Task-Oriented Approach to Malfunction Diagnosis in Complex Processing Plants," *NSF-AAAI Workshop on AI in Process Engineering*, Columbia University, New York (Mar. 9-10, 1987).

Denn, M. M., *Process Fluid Mechanics*, Prentice-Hall, Englewood Cliffs, NJ (1980).

Himmelblau, D. M., *Fault Detection and Diagnosis in Chemical and Petrochemical Processes*, p. 44, Elsevier, Amsterdam (1978).

Hoskins, J. C., K. M. Kaliyur, and D. M. Himmelblau, "Fault Diagnosis in Complex Chemical Plants Using Artificial Neural Networks," *AIChE J.*, **37**(12), 137 (1991).

Isermann, R., "Process Fault Detection Based on Modeling and Estimation Methods—A Survey," *Automatica*, **20**, 387 (1984).

Kramer, M. A., and B. L. Palowitch, Jr., "A Rule-Based Approach to Fault Diagnosis Using the Signed Directed Graph," *AIChE J.*, **33**(7), 1067 (1987).

Lamb, D. E., P. Dhurjati, and D. L. Chester, "Development of an Expert System for Fault Identification in a Commercial Scale Chemical Process," *NSF-AAAI Workshop on AI in Process Engineering*, Columbia University, New York (Mar. 9-10, 1987).

Lees, F. P., "Process Computer Alarm and Disturbance Analysis: Review of the State of the Art," *Comput. Chem. Eng.*, **7**(6), 669 (1983).

Li, R., and J. H. Olson, "Fault Detection and Diagnosis in a Closed-Loop Nonlinear Distillation Process: Application of Extended Kalman Filters," *Ind. Eng. Chem. Res.*, **30**, 898 (1991).

Ljung, L., *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ (1987).

Mah, K. N., *Design of Fault Monitoring Systems in Pipeline Networks*, MS Thesis, National Cheng Kung University, Tainan, Taiwan (1992).

Park, S., and D. M. Himmelblau, "Fault Detection and Diagnosis via Parameter Estimation in Lumped Dynamic Systems," *Ind. Eng. Chem. Process Des. Dev.*, **22**, 482 (1983).

Petti, T. F., J. Klein, and P. S. Dhurjati, "Diagnostic Model Processor: Using Deep Knowledge for Process Fault Diagnosis," *AIChE J.*, **36**(4), 565 (1990).

Rich, S. H., and V. Venkatasubramanian, "Model-Based Reasoning in Diagnostic Expert Systems for Chemical Process Plants," *Comput. Chem. Eng.*, **11**(2), 111 (1987).

Steward, D. V., "Partitioning and Tearing Systems of Equations," *SIAM J.*, **B2**(2), 345 (1965).

Venkatasubramanian, V., and K. Chan, "A Neural Network Methodology for Process Fault Diagnosis," *AIChE J.*, **35**(11), 1993 (1989).

Venkatasubramanian, V., R. Vaidyanathan, and Y. Yamamoto, "Process Fault Detection and Diagnosis Using Neural Network—I. Steady-State Processes," *Comput. Chem. Eng.*, **14**, 699 (1990).

Watanabe, K., and D. M. Himmelblau, "Fault Diagnosis in Nonlinear Chemical Processes, Part I. Theory," *AIChE J.*, **29**(2), 243 (1983a).

Watanabe, K., and D. M. Himmelblau, "Fault Diagnosis in Nonlinear Chemical Processes, Part II. Application to a Chemical Reactor," *AIChE J.*, **29**(2), 250 (1983b).

Watanabe, K., and D. M. Himmelblau, "Incipient Fault Diagnosis of Nonlinear Processes with Multiple Causes of Faults," *Chem. Eng. Sci.*, **39**(3), 491 (1984).

Watanabe, K., Matsuura M. Abe, M. Kubota, and D. M. Himmelblau, "Incipient Fault Diagnosis of Chemical Process via Artificial Neural Networks," *AIChE J.*, **35**(11), 1803 (1989).

Willsky, A. S., "A Survey of Design Methods for Failure Detection in Dynamic Systems," *Automatica*, **12**, 601 (1976).

Yu, C.-C., and C. Lee, "Fault Diagnosis Based on Qualitative/Quantitative Process Knowledge," *AIChE J.*, **37**(4), 617 (1991).

Appendix

For the completeness of this article, a modified version of the partitioning algorithm suggested by Steward (1965) is presented here. Notice that instead of simultaneous algebraic equations, a system of ordinary differential equations is the subject of our analysis. Thus, the output variable corresponding to each equation is automatically determined and there is no need for selecting the output set, which is a critical step in Steward's algorithm. After constructing the structural matrix, the following procedure can be followed to obtain a partition of the system:

1. We look for a row with no off-diagonal element and eliminate that row and the column corresponding to it. We repeat this process until there are no further rows without off-diagonal elements.

2. We begin tracing a path through the structural matrix by following the off-diagonal elements in search of a loop as follows:

(a) Select the first row remaining in the matrix as the "row to be examined" and enter its row number on a list.
(b) Locate the first off-diagonal element in the row being examined.
(c) Select the row corresponding to the column in which the off-diagonal element was found as the next row to be examined and add the row number to the list of rows examined.
(d) If the new row number has not previously been examined (that is, is not already on the list), return to Step b and continue tracing.
(e) If the new row number is already on the list, then we have found a loop containing all of the rows whose numbers appear on the list between the two occurrences of the last row number on the list.

3. When we find a loop, we replace the set of rows in the loop by one row which is the *union* of the rows replaced. The union of the two rows is a row which contains an element in each column in which *either* row originally contained an element. This we call *collapsing* the rows in the loop. Similarly, we collapse the columns corresponding to these rows.

4. We proceed to Step 1 and look for a row with no off-diagonal element. When a row is eliminated in Step 1, that row and the rows which collapse to form it represent the equations in a block. The order in which rows without off-diagonal elements are eliminated gives an order in which the changes in the variables of these blocks may occur.

Manuscript received Aug. 12, 1992, and revision received Dec. 21, 1992.

See NAPS document no. 05034 for 8 pages of supplementary material. Order from NAPS c/o Microfiche Publications, P.O. Box 3513, Grand Central Station, New York, NY 10163. Remit in advance in U.S. funds only $7.75 for photocopies or $4.00 for microfiche. Outside the U.S. and Canada, add postage of $4.50 for the first 20 pages and $1.00 for each of 10 pages of material thereafter, $1.50 for microfiche postage.