Simplification Techniques for EKF Computations in Fault Diagnosis: Model Decomposition

Chuei-Tin Chang and Jung-Ing Hwang
Dept. of Chemical Engineering, National Cheng Kung University, Tainan, Taiwan 70101, Republic of China

The extended Kalman filter (EKF) is one of the most popular model-based techniques for fault detection and diagnosis. In this study, the suboptimal EKF technique is utilized to enhance computation efficiency without sacrificing diagnostic accuracy. In particular, three simple strategies are proposed to decompose the filter model according to the precedence order of the state/parameter estimation process. The computation load needed in fault identification can be reduced significantly by implementing all or part of these decomposed EKFs on-line. Extensive simulation results are also presented to demonstrate the effectiveness of these proposed techniques.

Introduction

Due to the frequency and seriousness of chemical accidents that have occured in recent years, the importance of incipient fault detection and diagnosis in complex process plants has become apparent. Among various different model-based approaches adopted in the past, the extended Kalman filter (EKF) is clearly one of the most popular methods, (see Watanabe and Himmelblau, 1983a,b, 1984). In essence, EKFs of one form or another were employed to estimate both the states and parameters of chemical engineering systems, and then causes of abnormal system behaviors were identified accordingly.

Although the effectiveness of the EKF has been widely recognized, its use in commercial units has been, in fact, very limited. This is mainly due to a critical drawback of EKF, namely, its inability to guarantee unbiased estimates (Watanabe and Himmelblau, 1984). Obviously, incorrect information about the system parameters and/or states can mislead diagnosis. To overcome this problem, we modified the traditional way of implementing the EKFs (Chang et al., 1993; Chang and Chen, 1995). Instead of estimating all parameters simultaneously in a large EKF, several EKFs were used in parallel. Although this approach served the purpose of eliminating bias and misdiagnosis, the computational effort needed to carry out the parallel-parameter estimation scheme can be overwhelming.

The present article is the first of two companion articles that address this practical issue. Basically, the idea of a suboptimal Kalman filter was utilized in our studies to enhance computation efficiency without sacrificing diagnostic performance. There are in general two approaches to achieve this purpose: choosing simplified system models, and choosing simplified filter gains (Gelb, 1974). For illustration convenience, the latter approach is discussed in a separate article. This article is primarily concerned with the former, that is, replacing the system model with simpler ones in the EKF computations. Notice that such a practice has never been systematically studied before.

In order to facilitate understanding of the proposed simplification strategies, some background information is provided in the next three sections. The key concept of *fault observability* (Chang and Chen, 1995) is reviewed first to avoid confusion. The pattern of estimated propagation in EKF computations is then analyzed in detail. New data concerning the relative magnitudes of the adjustments in updating the estimates of parameters and states are also presented to justify a critical assumption, namely, that corrections needed in state updates are usually negligible. Finally, the algorithms for establishing *precedence order of influences* among state variables and then identifying fault-observable system structures are described with several examples.

The rest of this article is concerned with a new simplification procedure for applying EKFs in fault diagnosis. Specifically, the original EKF is decomposed into several suboptimal but smaller EKFs according to the precedence order. This task can be easily accomplished with three simple decoupling techniques developed in this study. The computation demand can then be greatly reduced by adopting these smaller EKFs

Correspondence concerning this article should be addressed to C.-T. Chang.

on-line. In addition, the computation process can be simplified even further by judiciously removing some of the EKFs that do not produce estimates of interest.

Fault Observability

The EFK is used in this application to estimate both states and *parameters* of the system model. This is because parameter estimates are in general more sensitive to faults than estimates of the state variables, and thus they are better indications of the degradation of system performance (Dalle Molle and Himmelblau, 1987). Since it is usually possible to associate the assumed malfunctions with changes in the corresponding model parameters, these parameters can be treated as augmented states in the corresponding EKF (Himmelblau, 1978). Specifically, let us consider a system model with the following general form:

$$\frac{dx}{dt} = f(x, \theta, t) + \omega_x; \qquad \omega_x \sim \mathfrak{N}(\mathbf{0}, Q), \tag{1}$$

in which

$$f(\cdot) = [f_1(\cdot), f_2(\cdot), \cdots, f_n(\cdot)]^T,$$

$$x = [x_1, x_2, \cdots, x_n]^T \qquad \theta = [\theta_1, \theta_2, \cdots, \theta_m]^T$$

$$\omega_x = [\omega_1, \omega_2, \cdots, \omega_n]^T,$$

where x_i represents the state variables and θ_j the parameters or inputs of the system; f_k represents nonlinear functions of x_i and θ_j, ω_k represents the normally distributed random system noises, and Q is the covariance matrix associated with ω_x.

In order to estimate the time-variant parameters and/or inputs (θ_j) in an EKF, one can treat them as state variables and augment the corresponding equations with Eq. 2,

$$\frac{d\tilde{x}}{dt} = \frac{d}{dt}\begin{bmatrix} x \\ \theta \end{bmatrix} = \begin{bmatrix} f \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \omega_x \\ \omega_\theta \end{bmatrix} = \tilde{f} + \tilde{\omega}; \qquad \tilde{\omega} \sim \mathfrak{N}(\mathbf{0}, \tilde{Q}), \tag{2}$$

where ω_θ is an m-dimensional random vector with mean equal to zero. For the sake of convenience, the components in $\tilde{\omega}$ are assumed to be independent, and thus $\tilde{Q}$ is a diagonal matrix. Also, without loss of generality, it is assumed in this study that the first s ($s \leq n$) state variables can be measured directly. In other words, the measurement model can be written as

$$z_l = H\tilde{x}_l + v_l = [I \mid \mathbf{0}]\tilde{x}_l + v_l; \qquad v_l \sim \mathfrak{N}(\mathbf{0}, R), \tag{3}$$

where z_l, $\tilde{x}_l$, and v_l are the system output vector, augmented state vector, and measurement noise vector, respectively, at time t_l; I is an $s \times s$ identity matrix; and $\mathbf{0}$ is an $s \times (m + n - s)$ matrix whose entries are all zeroes. Also, R is assumed to be a diagonal matrix in this study.

One of the obvious reasons for the failure of a Kalman filter to produce unbiased estimates is that the system itself is *unobservable* (Grewal and Andrews, 1993). However, the traditional criteria for testing system observability cannot be used as sufficient conditions to guarantee the correctness of parameter estimates (Chang and Chen, 1995). It is thus highly desirable to develop a systematic approach for identifying EKFs that always produce accurate estimates if their models are correct. Such filters will be referred to as *fault observable* EKFs in this article.

Estimate Propagation

To develop a systematic method for testing fault observability, it is obvious that a thorough understanding of the estimation algorithm is necessary. First of all, notice that it is a common practice to adopt a diagonal system noise covariance matrix, that is, $\tilde{Q}$ in Eq. 2, in which the variances associated with the augmented parameters are much larger than those with the states (Watanabe and Himmelblau, 1984). This is due primarily to the belief that the model parameters are more sensitive than the state variables to incipient faults. As a result, the corrections in the updated estimates of the state variables are usually negligible when compared with those of the parameters. In other words, the *relative magnitude of adjustment* (RMA) in parameter must be much larger than that in state:

$$\frac{\hat{\theta}_i^{(+)}(t) - \hat{\theta}_i^{(-)}(t)}{\hat{\theta}_i^{(+)}(t)} \gg \frac{\hat{x}_j^{(+)}(t) - \hat{x}_j^{(-)}(t)}{\hat{x}_j^{(+)}(t)}$$

$$i = 1, 2, \cdots, m \qquad j = 1, 2, \cdots, n, \tag{4}$$

where $\hat{\theta}_i^{(-)}(t)$ and $\hat{x}_j^{(-)}(t)$ represent the estimates of model parameters and state variables obtained by integrating Eqs. 2: $\hat{\theta}_i^{(+)}(t)$ and $\hat{x}_j^{(+)}(t)$ denote the corresponding updated estimates obtained on the basis of online measurements. This phenomenon can be demonstrated in the following example:

Example 1. Let us consider the system of two identical storage tanks connected in series (see Figure 1). The system model can be written as

$$A_1 \frac{dh_1}{dt} = q_i - q_1 \tag{5}$$

$$\frac{dq_1}{dt} = \frac{\pi d_1^2}{4\rho l_1}\left[\rho g(h_1 - h_2) - \frac{8(f_1 + \Delta f_1)l_1 \rho q_1 |q_1|}{\pi^2 d_1^5}\right] \tag{6}$$

$$A_2 \frac{dh_2}{dt} = q_1 - q_2 \tag{7}$$

$$\frac{dq_2}{dt} = \frac{\pi d_2^2}{4\rho l_2}\left[\rho g h_2 - \frac{8 f_2 l_2 \rho q_2 |q_2|}{\pi^2 d_2^5}\right], \tag{8}$$

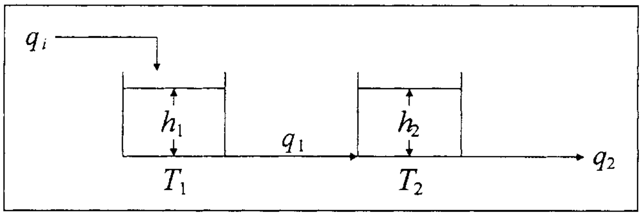

Figure 1. Simplified process flow diagram of a two-tank system.

where ρ ($=1{,}000.0$ kg/m^3) is the density of liquid; h_k and A_k ($=1.0$ m^2) denote, respectively, the height of liquid level in and the cross-sectional area of tank k ($k=1$, 2); q_k, d_k ($=0.0508$ m), l_k ($=5.0$ m) and f_k ($=2.509\times 10^{-3}$) represent, respectively, the volumetric flow rate in and the diameter, length, and friction factor of the outlet pipeline from tank k ($k=1$, 2). Also notice that the parameter Δf_1 is associated with the assumed failure, that is, partial blockage in the pipeline between tank 1 and tank 2.

Numerical simulation studies were carried out to verify the correctness of Eq. 4. It was assumed that the system was operated at its normal steady state initially. The initial heights of liquid levels h_1 and h_2 were chosen to be 1.378 m and 0.689 m, respectively, and the corresponding flow rates were $q_i = q_1 = q_2 = 0.015$ m^3/s. The fault just mentioned occurred at $t=50$ s. More specifically, the change in Δf_1 was decribed as

$$\Delta f_1 = C_f\{1-\exp-\alpha(t-50)\}u(t-50) \tag{9}$$

and

$$u(t-50)=\begin{cases}1 & t\geq 50\\ 0 & t<50\end{cases}, \tag{10}$$

where C_f ($=0.001$) and α ($=0.05$ s^{-1}) are constants.

Equations 5–10 were integrated together to produce the transient behavior of the state variables. In this example, it was further assumed that the state variables h_1 and q_2 can be measured on-line. The measurement noises were produced with a random-number generator and then added to the simulated values of these two variables to obtain the simulated on-line measurements.

The corresponding EKF was then applied to the simulated measurement data. The covariance matrix $\tilde{Q}$ adopted in this EKF was of the form:

$$\tilde{Q}=\begin{bmatrix}0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&Q_5\end{bmatrix}. \tag{11}$$

From the results of extensive simulation studies, we found that it is always possible to obtain correct estimates of both the parameter and states. A sample of the parameter estimates is presented in Figure 2. In addition, the corresponding RMAs were computed and plotted (Figure 3). From these results, one can see that the RMA in Δf_1, that is, $r_{\Delta f1}$, is indeed much larger that those in states, that is, r_{h_1}, r_{q_1}, r_{h_2}, and r_{q_2}, even after the initialization period.

If an EKF is to be used for fault detection and diagnosis, its state-estimate propagation equations must be formulated according to Eq. 2. At any sampling time (say t_{l-1}), these equations should be integrated numerically to determine the estimates at the next time step, t_l, based on the update estimates at t_{l-1}. Whichever numerical method is used in this application, a set of nonlinear algebraic equations should always be solved simultaneously. As mentioned previously, one can neglect the adjustments in states,

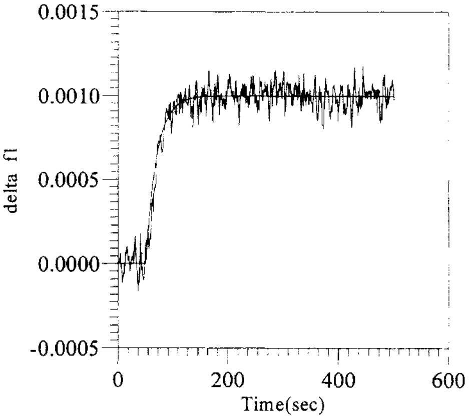

Figure 2. EKF estimates of Δf_1 in Example 1.

$$\hat{x}_i^{(+)}(t_{l-1}) \sim \hat{x}_i^{(-)}(t_{l-1}). \tag{12}$$

Also from Eq. 2

$$\hat{\theta}_i^{(+)}(t_{l-1}) = \hat{\theta}_i^{(-)}(t_l). \tag{13}$$

Therefore, the task of numerically integrating the state-estimate propagation equations is essentially one of solving

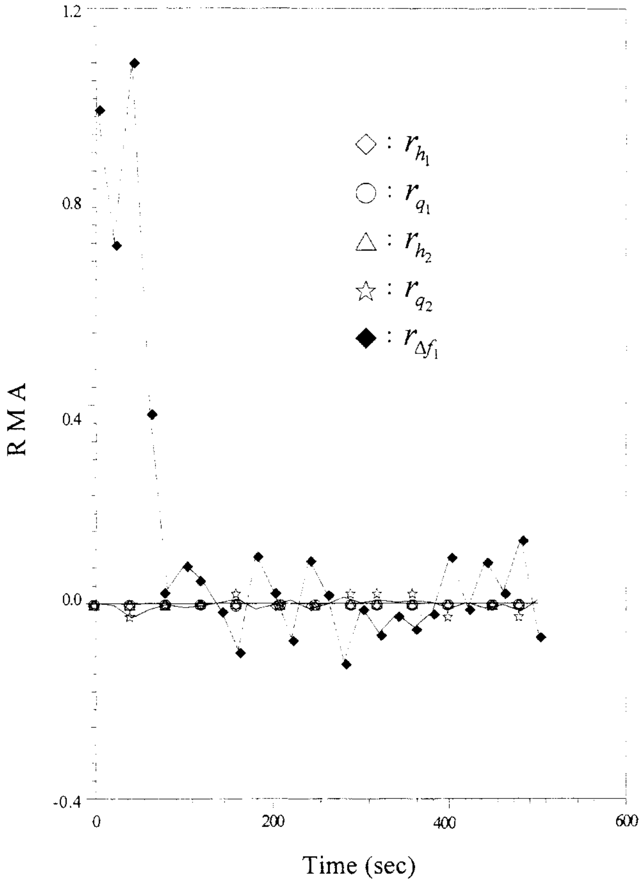

Figure 3. Relative magnitude of adjustment in Δf_1, h_1, q_1, h_2, and q_2.

simultaneous algebraic equations of the following general form:

$$\phi_1[\hat{x}_1^{(-)}(t_l), \cdots, \hat{x}_n^{(-)}(t_l); \hat{\theta}_1^{(+)}(t_{l-1}), \cdots, \hat{\theta}_n^{(+)}(t_{l-1})] = 0$$

$$\phi_n[\hat{x}_1^{(-)}(t_l), \cdots, \hat{x}_n^{(-)}(t_l); \hat{\theta}_1^{(+)}(t_{l-1}), \cdots, \hat{\theta}_n^{(+)}(t_{l-1})] = 0, \tag{14}$$

where $\hat{x}_i^{(-)}(t_l)$ $(i = 1, 2, \cdots, n)$ are the estimates of states at time t_l, which should be considered as the unknowns of Eqs. 14. On the other hand, $\hat{\theta}_j^{(+)}(t_{l-1})$ are the *updated* estimates of parameters at time t_{l-1}. At time t_{l-1}, one first needs to determine the values of $\hat{\theta}_j^{(+)}(t_{l-1})$ and then Eqs. 14 can be solved accordingly. In other words, the state estimates at time t_l are mainly dependent upon the updated estimates of the model parameters at time t_{l-1}.

If the EKF performs satisfactorily, correct estimates of $\hat{\theta}_j^{(+)}(t_{l-1})$ can be chosen to yield the following results:

$$\hat{x}_i^{(-)}(t_l) \sim \bar{x}_i \qquad i = 1, 2, \cdots, s, \tag{15}$$

where $\bar{x}_i$ represents the measurement values of the state variables at time t_l. However, it is also a well-known fact that incorrect estimation is a phenomenon often encountered in the practical applications of EKF. Thus, the structure of the state-estimate propagation equations, that is, Eqs. 14, must be further analyzed to gain additional insights for identifying all possible causes of estimation bias. Specifically, the *precedence order of influences* among the state variables $\hat{x}_i^{(-)}(t_l)$ must be established for this purpose due to changes in the adjustable parameters $\hat{\theta}_j^{(+)}(t_{l-1})$.

Structural Analysis

The main thrust of structural analysis is to identify fault observable EKFs with simple qualitative techniques. Specifically, the partition algorithm suggested by Steward (1965) is used in this work as the basic procedure to determine the precedence order. This original algorithm will be referred to as *Algorithm A*. For the sake of illustration convenience, it is also included in the Appendix.

To facilitate understanding the rationale behind various steps in the proposed procedure for testing fault observability, a series of examples is presented below. First of all, it is intuitively correct that, in a fault observable system, the symptoms of faults must appear in the measurement data. Precedence order can be used as an aid to determine whether this criterion is satisfied. This fact can be demonstrated with a simple example.

Example 2. Let us consider the two-tank system presented in Figure 4. The system model can be written as

$$A_i \frac{dh_1}{dt} = q_i - q_1 - \Delta cl_1 \sqrt{h_1} \tag{16}$$

$$\frac{dq_1}{dt} = \frac{\pi d_1^2}{4 \rho l_1} \left[\rho g h_1 - \frac{8 f_1 l_1 \rho q_1 | q_1 |}{\pi^2 d_1^5} \right] \tag{17}$$

$$A_2 \frac{dh_2}{dt} = q_1 - q_2 \tag{18}$$

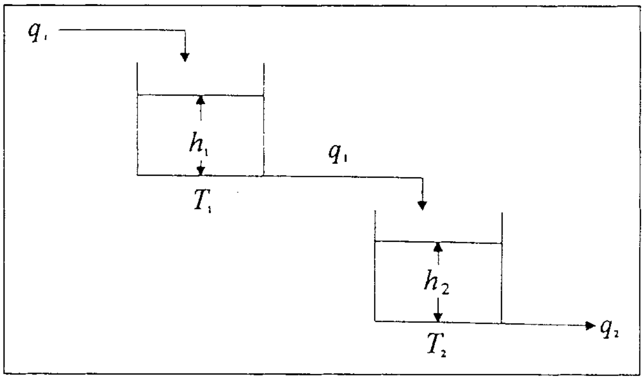

Figure 4. Simplified process flow diagram of another two-tank system.

$$\frac{dq_2}{dt} = \frac{\pi d_2^2}{4 \rho l_2} \left[\rho g h_2 - \frac{8 (f_2 + \Delta f_2) l_2 \rho q_2 | q_2 |}{\pi^2 d_2^5} \right], \tag{19}$$

where the parameters Δcl_1 and Δf_2 are associated with two assumed failures, namely, leakage in tank 1 and partial blockage in the exit pipeline of tank 2, respectively. As indicated previously, the state-estimate propagation equations (Eqs. 14) can be solved if the values of the model parameters are given. The corresponding precedence order can be determined with Algorithm A. The result is represented with a precedence diagram (Figure 5). From Figure 5 one can see that all four state variables in this system are affected by a leak in tank 1, but partial blockage in pipeline 2 can only cause h_2 and q_2 to behave abnormally. Thus, if h_1 and q_1 (or only one of them) are chosen as the measurement variables, it is certainly not possible to produce correct estimates of Δf_2 on the basis of the available on-line data.

Although application of Algorithm A yields a precedence order that is useful for identifying a special class of diagnostically unobservable systems, this approach is still limited in the sense that the correctness and uniqueness of the solutions to the state estimate propagation equations cannot be confirmed accordingly. This situation can be illustrated with another example.

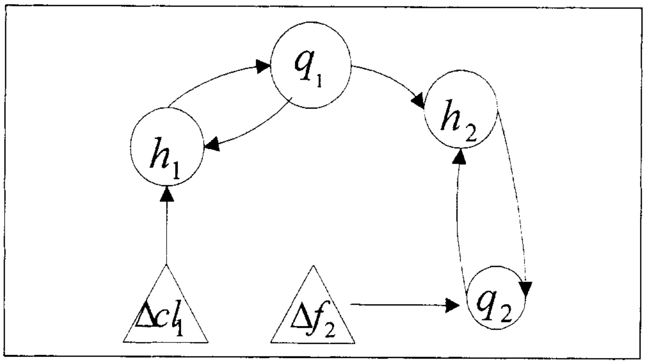

Figure 5. Precedence diagram of the two-tank system in Figure 4.

Result of implementing Algorithm A (fault parameters: Δcl_1 and Δf_2).

Example 3. Let us again consider the system presented in Figure 1. It is assumed that there are two possible faults: (1) a sudden change in the inlet flow rate, and (2) partial blockage in the pipeline between tank T_1 and tank T_2. Again, Algorithm A has been applied and the precedence diagram can be obtained accordingly (see Figure 6). Notice that Δq_i and Δf_1 are the parameters associated with faults (1) and (2), respectively. From Figure 6, it can be observed that all four state variables—h_1, q_1, h_2, and q_2—are affected by the two parameters Δq_i and Δf_1. Thus, if at least one state variable can be measured on-line, the occurrence of faults should be detectable. However, one can also find in the precedence diagram that all state variables are connected to *both* parameters, and they are interconnected by several feedback loops. In other words, all of them are within one block, and thus must be solved simultaneously. Therefore, on the basis of this precedence order, one still cannot be certain whether a *unique* set of correct parameter values can be found to satisfy the requirements implied in Eqs. 15.

If the measurement variables are embedded in coupled feedback loops, one can see from the preceding example that the precedence order obtained with the traditional approach is not really useful for the purpose of confirming fault observability. Thus, additional tools must be developed for our purpose.

It should be noted that it may not be necessary to guess and iterate all variables in solving an irreducible set of equations. For sparse equation sets, it is often possible to reach a solution by guessing only a few of the variables. This is the so-called *tearing* technique (Stadtherr et al., 1974), which can be used to determine an efficient iteration procedure. In this research, this method was also adopted as an aid for clarifying the cause-and-effect relations between the model parameters and measurement variables in the state-estimate propagation equations. In particular, all s measurement variables were treated as the "tear variables" (the variables that are guessed), and s "tear equations"(the equations used to check the guesses) were then chosen from Eqs. 14. These tear equations were chosen with a simple criterion, namely, each equation must contain the corresponding tear variable. The tearing operation can be performed on the precedence diagram obtained with Algorithm A. Specifically, the edges between the tear variables and their outputs are removed from the original digraph. In this article, the term *Algorithm B* will be used to indicate the procedure of implementing Algorithm A after the proposed tearing steps. The advantage of Algorithm B can be demonstrated clearly with the following example.

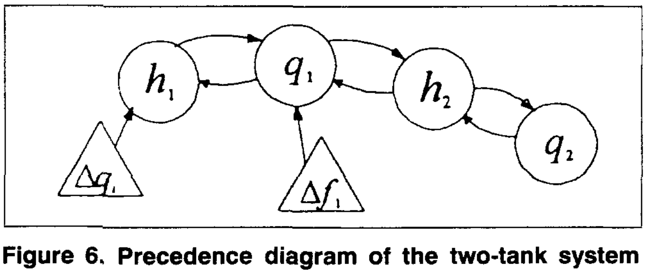

Figure 6. Precedence diagram of the two-tank system in Figure 1.

Result of implementing Algorithm A (fault parameters: Δq_i and Δf_1).

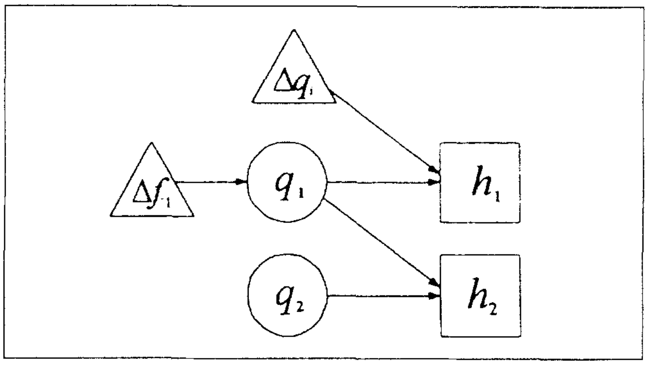

Figure 7. Precedence diagram of the two-tank system in Figure 1.

Result of implementing Algorithm B with h_1 and h_2 as the measurement variables (fault parameters: Δq_i and Δf_1).

Example 4. Let us reconsider the system described in Example 3. Assume that h_1 and h_2 are the measurement variables in this case, and thus should be regarded as the tear variables in the structural analysis. The tearing operation can be performed on Figure 6, and the resulting precedence order is presented in Figure 7. In the conventional process of solving algebraic equations, the tear variables are unknowns. Their values must be obtained through iterative procedure. In this work, however, the desired values of the measurement variables should satisfy the constraints stipulated in Eqs. 15. Thus, in solving the state-estimate propagation equations of this example, the values of tear variables can be set directly to be the measurement values, and the outputs h_1 and h_2 can then be calculated according to the precedence order in Figure 7 and any given values of Δf_1 and Δq_i. Also notice that, in this computation process, q_2 should be regarded as a constant since it is affected only by the tear variable h_2. Of course, the most appropriate parameter values should be chosen on the basis of Eqs. 15. From Figure 7, one can see that a change in either of the parameters Δf_1 and Δq_i can cause variations in one or both of the output variables, h_1 and h_2. Since the two parameters in this case must be adjusted *simultaneously* in order to produce the two desired output values, it is therefore assumed that the chance for biased EKF estimation in this situation is low and the system should be fault observable.

To facilitate later discussions, it is now necessary to classify the model parameters and measurement variables according to the precedence diagram just described. In particular, if a parameter is connected to one or more measurement variables, it is referred to as a "tunable parameter." On the other hand, if a measurement variable is connected to at least one parameter, then this variable is the "affected variable." If some of the tunable parameters can be assigned *independently* and the rest of the parameters can always be adjusted accordingly to produce the desired values for the affected variables, then there should be an infinite number of suitable parameter values that can satisfy Eqs. 15. As a result, the possibility of biased EKF estimation is extremely high and the system should also be regarded as diagnostically unobservable. The following example is used to demonstrate this special feature in EKF estimation.

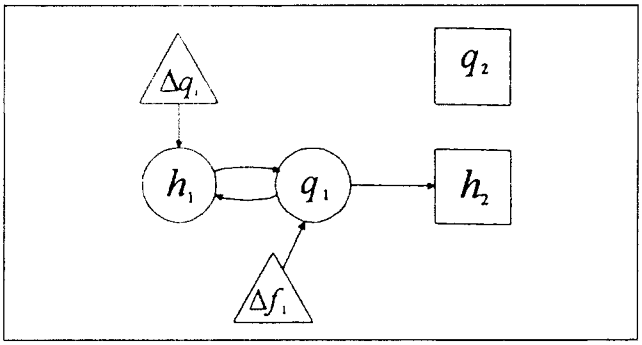

Figure 8. Precedence diagram of the two-tank system in Figure 1.

Result of implementing Algorithm B with h_2 and q_2 as the measurement variables (fault parameters: Δq_i and Δf_1).

Example 5. Let us again consider the system described in Example 3 and use h_2 and q_2 as the measurement variables this time. The results obtained with Algorithm B can be found in Figure 8. One can see that both Δq_i and Δf_1 are tunable parameters, but only h_2 is affected by these two parameters. Thus, the value of either one of the parameters can be assigned arbitrarily first and then the other parameter can always be adjusted to ensure output h_2 approaching its measurement value. Since the EKF does not have a prior knowledge about the actual variations in Δq_i and Δf_1, the possibility of obtaining the correct results is almost nil in the corresponding optimal estimation process.

Although the precedence order obtained after tearing can be adopted as the basis for identifying diagnostically unobservable systems, the methods mentioned before are still difficult to apply when the system is large and complicated. Thus, a systematic procedure was developed in an earlier study (Chang and Chen, 1995) to overcome this problem. It has to be emphasized that, although structural analysis is qualitative in nature and thus not theoretically rigorous, the correctness of its predictions has been verified in numerous simulation studies (Chen, 1993).

Decomposition Strategies

Applications of the techniques described earlier are in fact not limited to the development of tests for fault observability only. The insights gained with structural analysis offer clues to properly decompose the EKF model without sacrificing diagnostic performance. Three methods have been developed in this study:

Method 1

First of all, the size of the EKF model can be reduced according to the precedence diagram obtained with Algorithm A. In particular, several *blocks* can be identified. Since specific faults are assumed to occur in an EKF, it is only necessary to consider those blocks that are affected by the assumed faults, that is, the blocks in which the corresponding parameters appear and their downstream blocks. If some of the upstream variables are included in this subset of the model equations, their values should be considered to be at the normal levels without variations. This method is illustrated with the following example.

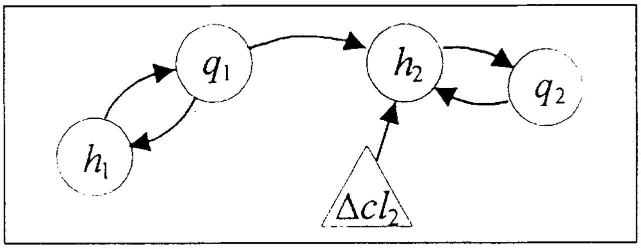

Figure 9. Precedence diagram of the two-tank system in Figure 4.

Result of implementing Algorithm A (fault parameter: Δcl_2).

Example 6. Let us consider the system in Figure 4 and assume that a leak may develop in the second tank. This fault is described with a parameter Δcl_2. Algorithm A can be applied to produce the precedence diagram in Figure 9. Since only the block (h_2, q_2) is affected by a change in Δcl_2, the upstream variable q_1 can be treated as a constant, that is, its initial steady-state value. As a result, the system in Figure 9 can be reduced to the one presented in Figure 10. The corresponding EKF model is

$$A_2 \frac{dh_2}{dt} = q_1^0 - q_2 - \Delta cl_2 \sqrt{h_2} \tag{20}$$

$$\frac{dp_2}{dt} = \frac{\pi d_2^2}{4\rho l_2}\left[\rho g h_2 - \frac{8 f_2 l_2 \rho q_2 | q_2 |}{\pi^2 d_2^5}\right] \tag{21}$$

$$\frac{d\Delta cl_2}{dt} = \omega_1(t), \tag{22}$$

where q_1^0 is the initial steady-state value of q_1. The feasibility of this approach has been verified with numerical simulation studies. The results are presented as Supplementary Material of this article (Figures S1 to S3).

Method 2

Secondly, an EKF may be decomposed into smaller ones by making use of the informations produced with Algorithm B. Notice that several separate *components* may be found in the resulting digraph. It is always possible to use an independent EKF to estimate the states and parameters within each

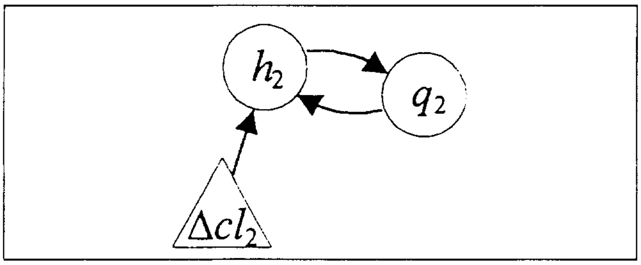

Figure 10. Results obtained by applying Method 1 to the precedence diagram in Figure 9.

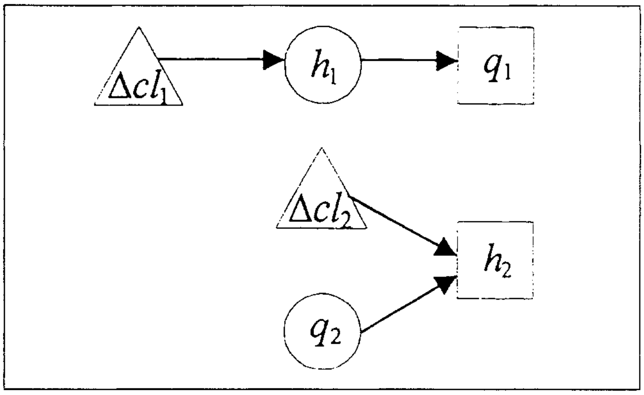

Figure 11. Precedence diagram of the two-tank system in Figure 1.

Result of implementing Algorithm B with q_1 and h_2 as the measurement variables (fault parameters: Δcl_1 and Δcl_2).

component. If the variables in other components appear in the model equations associated with a particular component, they must be the measured state variables and should be treated as the *augmented parameters* in the corresponding EKF. Let us use another example to illustrate this technique:

Example 7. Let us consider the system in Figure 1. Assume that there are two possible fault origins: leaks develop in tank T_1, or in tank T_2. Two parameters Δcl_1 and Δcl_2 are used to describe their effects. The measurement variables selected in this example are q_1 and h_2. The precedence diagram obtained with Algorithm B contains two components (see Figure 11). Consequently, two smaller EKFs can be adopted to estimate Δcl_1 and Δcl_2 on-line. Their respective models are

- EKF1

$$A_1 \frac{dh_1}{dt} = q_i - q_1 - \Delta cl_1 \sqrt{h_1} \tag{23}$$

$$\frac{dq_1}{dt} = \frac{\pi d_1^2}{4\rho l_1}\left[\rho g(h_1 - h_2) - \frac{8 f_1 l_1 \rho q_1 |q_1|}{\pi^2 d_1^5}\right] \tag{24}$$

$$\frac{dh_2}{dt} = \omega_1(t) \tag{25}$$

$$\frac{d\Delta cl_1}{dt} = \omega_2(t) \tag{26}$$

- EKF2

$$A_2 \frac{dh_2}{dt} = q_1 - q_2 - \Delta cl_2 \sqrt{h_2} \tag{27}$$

$$\frac{dq_2}{dt} = \frac{\pi d_2^2}{4\rho l_2}\left[\rho g h_2 - \frac{8 f_2 l_2 \rho q_2 |q_2|}{\pi^2 d_2^5}\right] \tag{28}$$

$$\frac{dq_1}{dt} = \omega_3(t) \tag{29}$$

$$\frac{d\Delta cl_2}{dt} = \omega_4(t). \tag{30}$$

These two EKFs have been tested again with numerically simulated data. The results can be found in the Supplementary Material (Figures S4 to S11).

Method 3

Finally, it should be noted that the digraph obtained with Algorithm B may contain only one component. In that case, it is still possible to divide this component into several subcomponents at some appropriately selected nodes that are associated with unmeasured state variables. These nodes are referred to as *severing nodes* in this work. Notice that the EKF used to estimate the states and parameters in any of these subcomponents must be smaller. In addition, if a subcomponent contains only states, the corresponding EKF computations can be omitted entirely.

A procedure for identifying the severing nodes follows:

1. Identify a candidate path that is initiated at a node without inputs and terminated at one corresponding to a measured variable.

2. Let the terminating node of the candidate path be the *current node* and check its upstream nodes on the path according to the following steps:

(a) Let the input of the current node on the candidate path be the *test node*.

(b) If the test node is not affected by any of the parameters, then go to the next step. Otherwise, let this test node be current node and repeat step (a).

(c) Treat the corresponding variable as an augmented parameter in EKF. In particular, other than its output edge on the candidate path, all the input and output edges of the test node are removed. The node symbol is then replaced with a Δ.

(d) Test fault observability of the resulting subcomponent. If this subcomponent is fault unobservable, then go to the next step. Otherwise, the test node should be considered as a severing node. Go to step 3.

(e) Treat the corresponding variable as a state in EKF. Specifically, recover the input edges that are removed in step 2(c) and change the node symbol back to ○.

(f) If the test node is not an initiating node, let the test node be the current node and then go to step 2(a). Otherwise, this initiating node should be regarded as a severing node.

3. Identify a different candidate path and repeat steps 2(a) to 2(f). This procedure is continued until all paths are exhausted.

An example is presented below to illustrate the preceding procedure and also demonstrate the effectiveness of the proposed approach.

Example 8. Let us reconsider Example 7 and use instead h_1 and h_2 as the measurement variables. The precedence diagram obtained with Algorithm B is presented in Figure 12. One can see clearly that in this case there is only one component and thus the size of corresponding EKF cannot be reduced with method 2.

Notice that three candidate paths can be identified in Figure 12: $q_1 \rightarrow h_1$, $q_1 \rightarrow h_2$, and $q_2 \rightarrow h_2$. Let us first examine the path $q_1 \rightarrow h_1$ and check q_1. If one treats the test node q_1 as an augmented parameter and removes all its edges that are not on the candidate path, the resulting subcomponent will be unobservable. Notice that q_1 is also an initiating node.

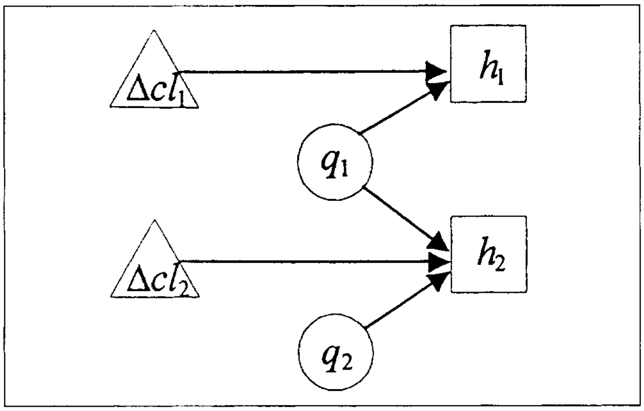

Figure 12. Precedence diagram of the two-tank system in Figure 1.

Result of implementing Algorithm B with h_1 and h_2 as the measurement variables (fault parameters: Δcl_1 and Δcl_2).

This node, therefore, should be viewed as a severing node and must be treated as a state in the EKF. The resulting subcomponent can be found in Figure 13. The corresponding EKF model should be the same as Eqs. 23–26. The effectiveness of this EKF has been confirmed with simulation studies and the results are presented in the Supplementary Material (see Figures S12–S15).

The preceding procedure can be applied to the path $q_1 \rightarrow h_2$. The conclusion of this exercise is similar, namely, q_2 is a severing node (see Figure 14) and should be treated as a state in EKF. The corresponding EKF model can be described with Eqs. 24, 27, 28, 30, and

$$\frac{dh_1}{dt} = \omega_3(t). \tag{31}$$

Again, simulation studies have been carried out to verify the correctness of this approach. The results are included in Figures S16 to S20 of the Supplementary Material.

Finally, since examination of the path $q_2 \rightarrow h_2$ does not result in a smaller subcomponent, no other alternative EKFs can be found with method 3.

Application Example

This example is designed to show the effectiveness of the proposed methods in lowering the computation load caused by applying EKFs on-line. The benefit of reducing the size of EKF is demonstrated here by counting the number of differential equations that are required to be integrated numerically. This number N_{eq} can be computed by

$$N_{eq} = \frac{(n+m)(n+m+1)}{2} + n. \tag{32}$$

Let us now consider the five-tank system presented in Figure 15 and assume that two parameters Δcl_4 and Δcl_5 are augmented in one of the EKFs used for diagnosis. From the precedence diagram produced with Algorithm A (Figure 16), one can see that Δcl_5 is located in the block corresponding to tank T_5 and Δcl_4 is in the block formed by T_3 and T_4. In other words, one only has to consider the precedence diagram presented in Figure 17. Consequently, the model equations associated with tanks T_1 and T_2 can be omitted and the variable q_2 should be set to a constant, $q_2 = q_2(0)$, in the resulting EKF. Specifically, the model used in this filter contains the following equations:

$$A_3 \frac{dh_3}{dt} = q_2 - q_4 - q_5 \tag{33}$$

$$A_4 \frac{dh_4}{dt} = q_5 - q_6 - \Delta cl_4 \sqrt{h_4} \tag{34}$$

$$A_5 \frac{dh_5}{dt} = q_4 - q_7 - \Delta cl_5 \sqrt{h_5} \tag{35}$$

$$\frac{dq_4}{dt} = \frac{\pi d_4^2}{4\rho l_4}\left[\rho g h_3 - \frac{8 f_4 l_4 \rho q_4 |q_4|}{\pi^2 d_4^5}\right] \tag{36}$$

$$\frac{dq_5}{dt} = \frac{\pi d_5^2}{4\rho l_5}\left[\rho g(h_3 - h_4) - \frac{8 f_5 l_5 \rho q_5 |q_5|}{\pi^2 d_5^5}\right] \tag{37}$$

$$\frac{dq_6}{dt} = \frac{\pi d_6^2}{4\rho l_6}\left[\rho g h_4 - \frac{8 f_6 l_6 \rho q_6 |q_6|}{\pi^2 d_6^5}\right] \tag{38}$$

$$\frac{dq_7}{dt} = \frac{\pi d_7^2}{4\rho l_7}\left[\rho g h_5 - \frac{8 f_7 l_7 \rho q_7 |q_7|}{\pi^2 d_7^5}\right] \tag{39}$$

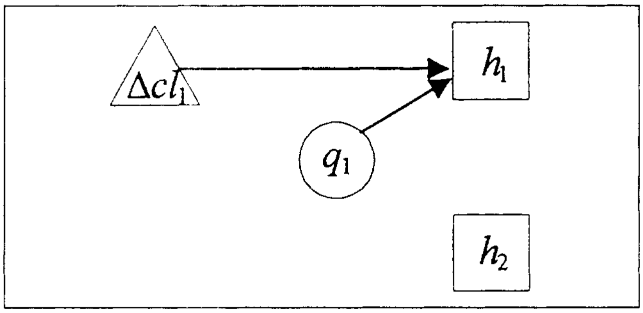

Figure 13. Subcomponent identified from the precedence diagram in Figure 12 with method 3.

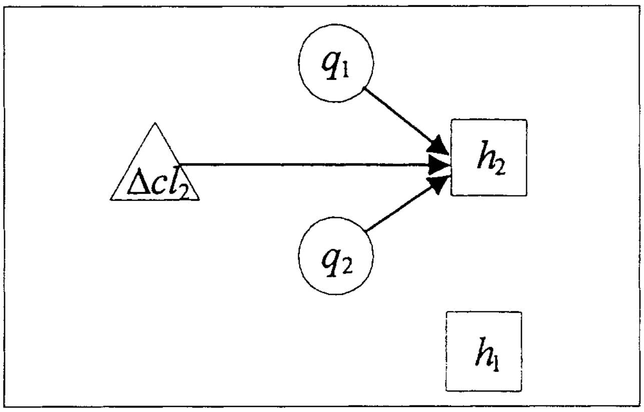

Figure 14. Another subcomponent identified from the precedence diagram in Figure 12 with method 3.

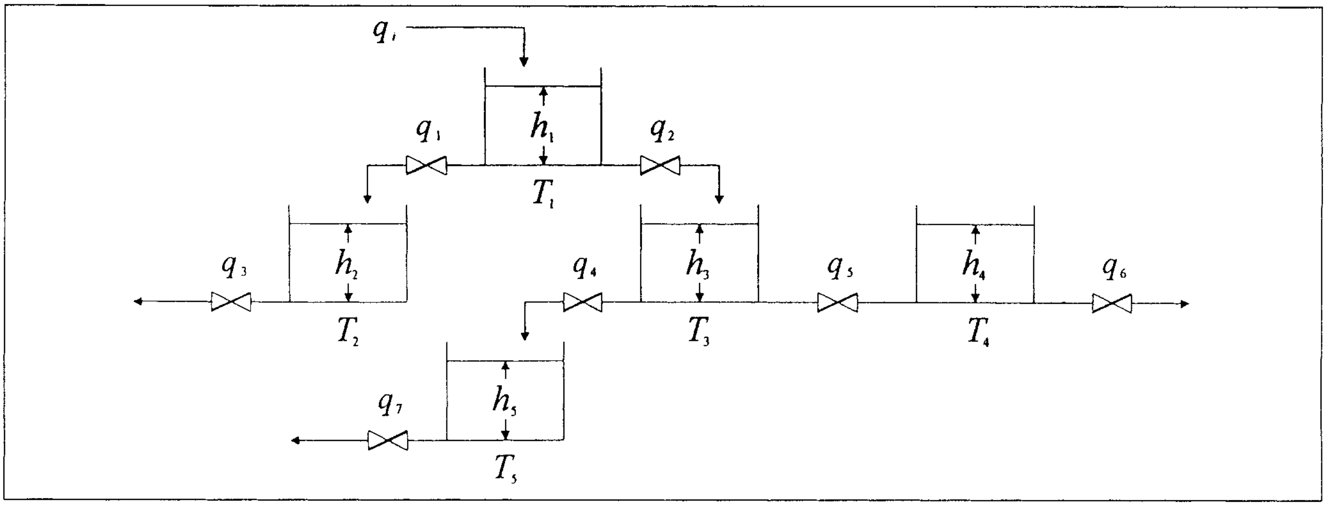

Figure 15. Simplified process flow diagram of a five-tank system.

$$\frac{d\Delta cl_4}{dt} = \omega_1(t) \tag{40}$$

$$\frac{d\Delta cl_5}{dt} = \omega_2(t). \tag{41}$$

The results of numerical simulation studies show that accurate estimates of states and parameters can be produced with this simplified EKF. On the basis of Eq. 32, one can see that the number N_{eq} is lowered significantly—from 117 to 52—by using method 1 alone.

Next, let us assume that five of the state variables—h_3, h_4, h_5, q_6, and q_7—are measured on-line. After applying Algorithm B on Figure 17, three separate components can be identified (Figure 18). Since the parameters Δcl_4 and Δcl_5 are located within one of the components, only one EKF is needed. In this EKF, three measured state variables—h_3, h_4, and h_5—and two unmeasured state variables—q_4 and q_5—are described with the original model equations, Eqs. 33-37. However, the other two measured variables—q_6 and q_7—and also Δcl_4 and Δcl_5 are all treated as the augmented parameters, as in Eqs. 40, 41, and

$$\frac{dq_6}{dt} = \omega_3(t) \tag{42}$$

$$\frac{dq_7}{dt} = \omega_4(t). \tag{43}$$

The number of equations requiring integration is thus reduced to 50. Again, numerical simulation studies have been carried out to verify the correctness of state and parameter estimation. The results are satisfactory. Notice that, in this case, the improvement in computation efficiency does not appear to be rewarding. This is due to the fact that a relatively large EKF is needed to estimate Δcl_4 and Δcl_5 with the present selection of the measured variables.

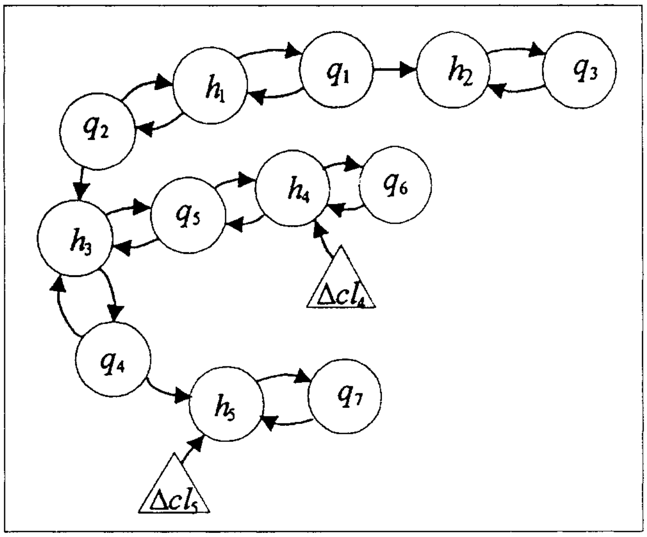

Figure 16. Precedence diagram of the five-tank system in Figure 15.

Result of implementing Algorithm A (fault parameters: Δcl_4 and Δcl_5).

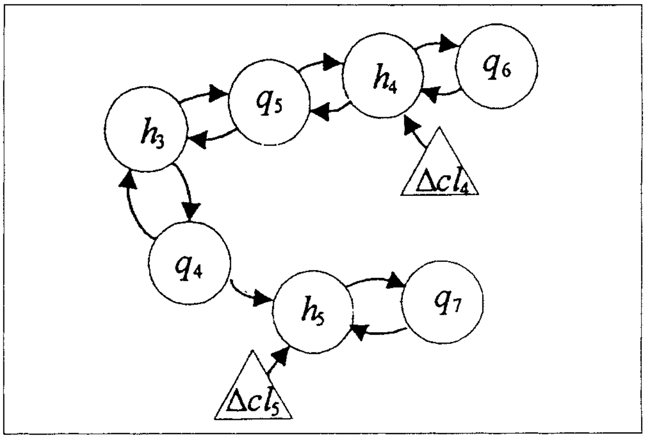

Figure 17. Result obtained by applying method 1 to the precedence diagram in Figure 16.

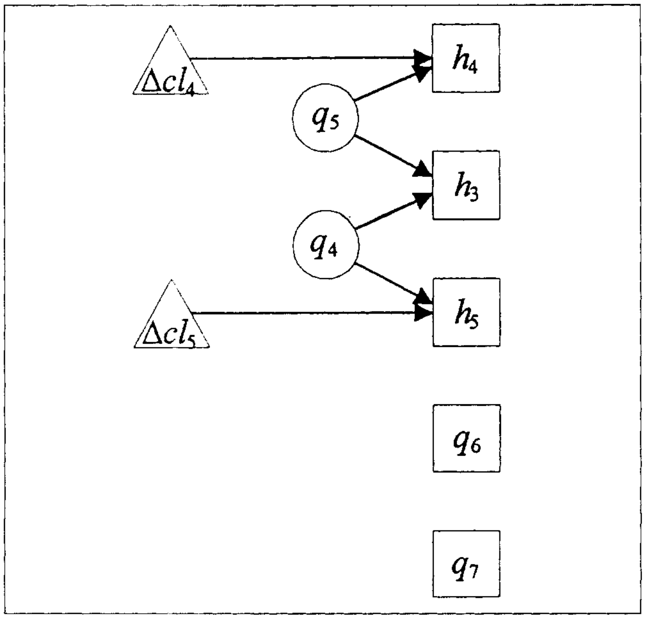

Figure 18. Precedence diagram obtained by applying Algorithm B to the subsystem represented by Figure 17.

With five measurement variables h_3, h_4, h_5, q_6, and q_7 (fault parameters: Δcl_1 and Δcl_2).

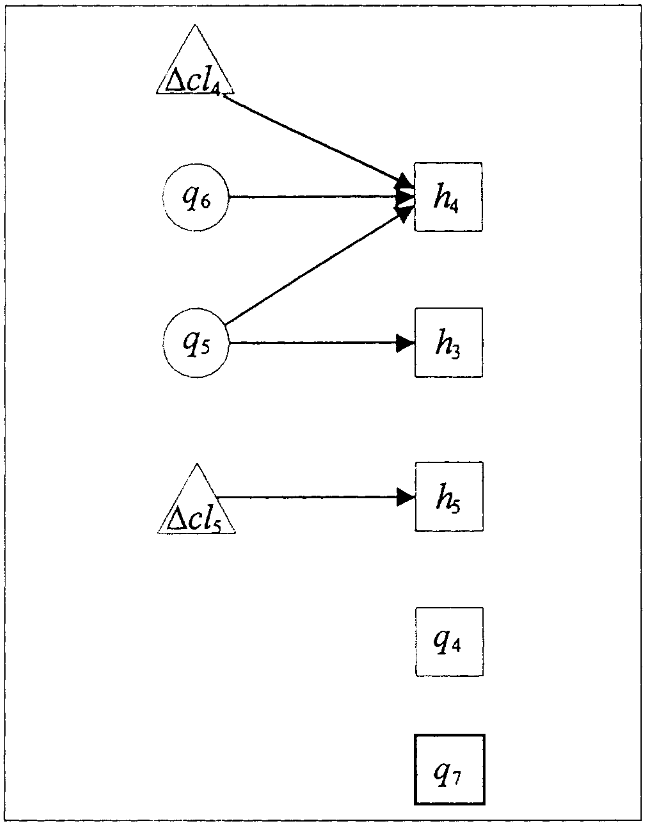

Figure 19. Precedence diagram obtained by applying Algorithm B to the subsystem represented by Figure 17.

With five measurement variables h_3, h_4, h_5, q_4, and q_7 (fault parameters: Δcl_1 and Δcl_2).

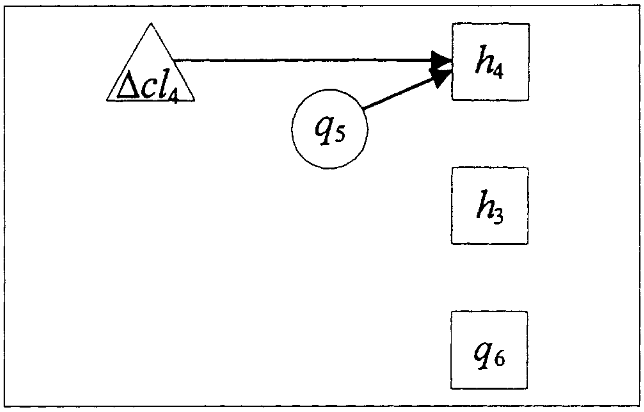

Figure 20. Subcomponent identified from the precedence diagram in Figure 18 with method 3.

If, for example, the measured variable q_6 is replaced by q_4, the precedence diagram can be divided into four components (see Figure 19). Since the parameters Δcl_4 and Δcl_5 are located in the first and second components, respectively, two smaller EKFs can be applied for diagnosis purpose. Specifically, the model corresponding to the first component can be described with Eqs. 33, 34, 37, 38, 40, and

$$\frac{dq_4}{dt} = \omega_5(t). \tag{44}$$

The model used in the second EKF consists of Eqs. 35, 41, 43, and 44. The feasibility of these two EKFs has been again verified with simulation studies. Further, it should be noted that the number N_{eq} can now be brought down to 36.

To illustrate the implementation procedure of method 3, let us reconsider the precedence diagram in Figure 18. Two severing nodes, q_4 and q_5, can be identified in the component containing Δcl_4 and Δcl_5. Consequently, two subcomponents can be obtained (see Figures 20 and 21). The EKF model associated with Figure 20 consists of Eqs. 34, 37, 40, 42, and

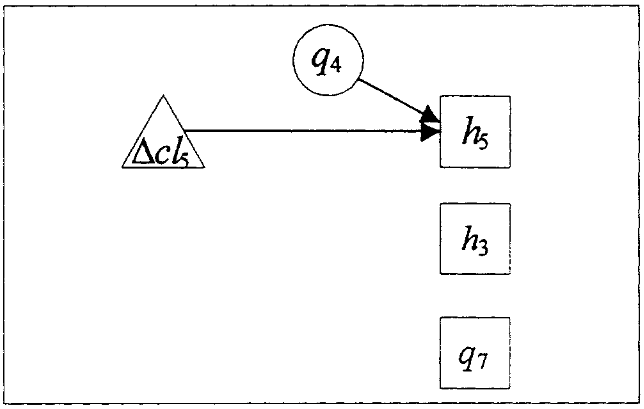

Figure 21. Another subcomponent identified from the precedence diagram in Figure 18 with method 3.

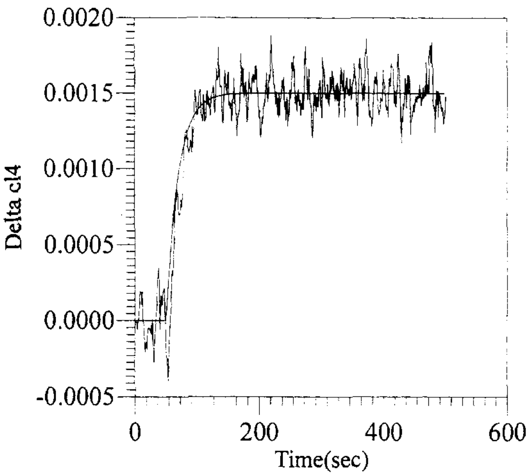

Figure 22. Estimates of Δcl_4 using the EKF corresponding to the subcomponent in Figure 20.

$$\frac{dh_3}{dt} = \omega_6(t). \tag{45}$$

On the other hand, the EKF model corresponding to the second subcomponent in Figure 21 can be described with Eqs. 35, 36, 41, 43, and 45. Notice that the number N_{eq} is now 17 for each EKF, and thus the total number of equations is 34. The effectiveness of these two EKFs has also been confirmed with numerical simulation. Samples of the results can be found in Figures 22 and 23, respectively.

Conclusions

A new approach has been proposed in this study to simplify EKF computations in fault identification without sacrificing diagnostic performance. In essence, this improvement is achieved with effective decoupling strategies that are developed on the basis of the precedence order of the state/parameter estimation process. From the results of extensive numerical simulation studies, one can see that the estimates of the simplified EKFs are indeed correct and, furthermore, the computation load can be reduced to less than 30% of the original level.

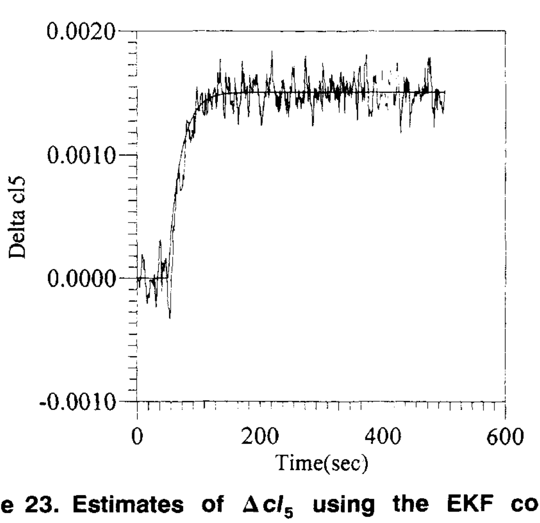

Figure 23. Estimates of Δcl_5 using the EKF corresponding to the subcomponent in Figure 21.

Acknowledgment

This work is supported by the National Science Council of the ROC under grant NSC85-2214-E006-015.

Literature Cited

Chang, C. T., and J. W. Chen, "Implementation Issues Concerning the EKF-Based Fault Diagnosis Techniques," *Chem. Eng. Sci.*, **50**(18), 2861 (1995).

Chang, C. T., K. N. Mah, and C. S. Tsai, "A Simple Design Strategy for Fault Monitoring Systems," *AIChE J.*, **39**(7), 1146 (1993).

Chen, J. W., *A Study on Fault Monitoring Techniques— Implementation Issues of the Extended Kalman Filters*, MS Thesis, National Cheng Kung University, Tainan, Taiwan, R.O.C (1993).

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.*, **84**(24), 2482 (1987).

Gelb, A., *Applied Optimal Estimation*, MIT Press, Cambridge, MA (1974).

Grewal, M. S., and A. P. Andrews, *Kalman Filtering-Theory and Practice*, Prentice Hall, Englewood Cliffs, NJ (1993).

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

Stadtherr, W. A., A. W. Gifford, and L. E. Scriven, "Efficient Solution of Sparse Sets of Design Equations," *Chem. Eng. Sci.*, **29**, 1025 (1974).

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

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).

Appendix: Algorithm A

For the completeness of this article, a modified version of the partitioning algorithm suggested by Steward (1965) is presented here. This algorithm can be best explained with the so-called *structural matrix*. Specifically, the (i, j)th entry in this array is filled with an × if the ith equation in Eqs. 14 involves the jth variable. Otherwise, it is blank. Notice that the diagonal positions are reserved for the output variables. One output must be chosen for each equation in Eqs. 14, and no variable becomes the output of more than one equation.

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

Table A1. Structural Matrix of the Two-Tank System in Example 2

	Δcl_1	Δf_2	h_1	q_1	h_2	q_2
ϕ_5	×					
ϕ_6		×				
ϕ_1	×		×	×		
ϕ_2			×	×		
ϕ_3				×	×	×
ϕ_4		×			×	×

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 (i.e., 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 that is the *union* of the rows replaced. The union of two rows is a row that 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 that collapsed 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 can occur.

As an example, notice that the precedence order in Figure 5 can be easily converted to the structural matrix presented in Table A1 and vice versa. The latter is actually the result of implementing Algorithm A.

Manuscript received July 28, 1997, and revision received Mar. 2, 1998.

See NAPS document no. 05460 for 22 pages of supplementary material. This is not a multiarticle document. Order from NAPS c/o Microfiche Publications, 248 Hempstead Turnpike, West Hempstead, NY 11552. Remit in advance in U.S. funds only $16.00 for photocopies or $5.00 for microfiche. There is a $25.00 invoicing charge on all orders filled before payment. 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, or $5.00 for the first microfiche and $1.00 for each fiche thereafter.