COMBINING AI, FDI, AND STATISTICAL HYPOTHESIS-TESTING IN A FRAMEWORK FOR DIAGNOSIS

Mattias Nyberg and Mattias Krysander

Dept. of Electrical Engineering, Linköping University,
SE-581 83 Linköping, Sweden
Phone: +46 13285714, Email: matny@isy.liu.se, matkr@isy.liu.se

Abstract: A new framework for model based diagnosis is presented using ideas from AI, FDI, and statistical hypothesis testing. The isolation mechanism is based on AI methods, and the main advantage is that multiple faults are handled implicitly. Thus, no special care for isolation of multiple faults is needed. The methods for residual generation, developed in the field of control theory (FDI), can within the framework be fully utilized. Since the framework is also based upon statistical hypothesis testing, it is suitable for problems including noise.

Keywords: fault diagnosis, AI-methods, fault isolation, FDI-methods, multiple faults, noise

1. INTRODUCTION

Fault diagnosis has in the literature been studied from mainly two different perspectives. The first is control theory (here denoted FDI), e.g. see (Gertler and Singer, 1990), and the second is AI, e.g. see (Kleer and Williams, 1987; Hamscher *et al.*, 1992). In the field of control theory, the literature on fault diagnosis has mostly been focused on the problem of *residual generation*. That is, given a model of the system, how to off-line construct residual signals that are zero in the fault-free case but sensitive to faults. In the field of AI, the focus has been on fault isolation and how to on-line compute what is here called residuals. In this paper we show how methods from FDI and AI (or more exactly *consistency based diagnosis*) can be combined into a common framework for fault diagnosis. The framework proposed is also based upon ideas from statistical hypothesis testing in accordance with the method *structured hypothesis tests* from (Nyberg, 1999; Nyberg, 2002).

The modelling of the system to be diagnosed, and the isolation of faults, follows mainly ideas from AI (Dressler *et al.*, 1993). The key point here is to add information in the model of how the validity of each model equation depends on which faults that are present in different components. Isolation is then performed by propagating this information through the diagnosis system by the use of standard logic. However, one difference is that residuals are computed off-line as in FDI. Therefore the on-line machinery can be made more simple, e.g. there is no need to use a so called ATMS (Assumption based Truth Maintenance System) which is common in AI (Kleer and Williams, 1987). All decisions taken in the diagnosis system are based on the theory of statistical hypothesis testing. This means for example that noise and uncertainties are handled in a sound way.

By combining these ideas from FDI, AI, and hypothesis testing, we will obtain a framework that is able to efficiently handle: fault models, several different fault types (e.g. parameter- and additive faults), more than two behavioral modes per component, general differential-algebraic models, noise, uncertainties, decoupling of disturbances, static and dynamic systems, and isolation of multiple faults.

The modelling framework and how information about different faults is incorporated in the model is described in Section 2. The design of a diagnosis system is then presented in Sections 3 and 4. The connection to FDI methods are more explicitly elaborated in Section 5. Section 6 describes how noise is treated, and finally, Section 7 discusses the output from the diagnosis system.

2. MODELLING FRAMEWORK

This section describes the modelling framework that is later used in the construction of the diagnosis system. Using this modelling framework, all information about the faults are included in the model. This fault information is then the basis for the reasoning about faults.

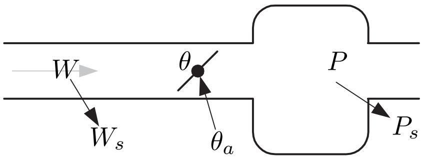

Fig. 1. An example of a gas-flow system.

Throughout the paper, we will exemplify all concepts and techniques on the same example. The example chosen is shown in Figure 1 and represents a gas-flow system. Gas flows into the system from the left. This flow is measured with a gas-flow sensor. The gas flow is controlled with a valve. Finally the gas pressure is measured with a pressure sensor. This kind of system is commonly found in for example combustion engines.

2.1 *Components*

We assume that the system consists of a number of components. The behavior of each component, and the relation to its outer world, are described by a number of relations.

A component has *internal* variables and *external* variables. External variables are variables that are shared with connected adjacent components[1] or can be observed. Internal variables are only known within the component itself.

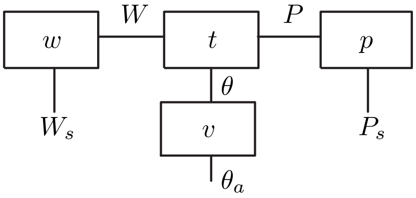

Fig. 2. Components of the gas-flow system.

The gas-flow system from Figure 1 can be separated into the components tube, gas-flow sensor, valve, and pressure sensor. These four components will be denoted t, w, v, and p respectively. The components and the connections between them are illustrated in Figure 2. In the figure it is seen that the tube t relates the variables θ, P, and W to each other. Further on, it is seen that the gas-flow sensor w relates the physical gas-flow W to the measured gas-flow W_s. Finally, the pressure sensor p relates the pressure P to the measured pressure P_s, and the valve v relates the physical angle θ to the actuated (i.e. demanded by the control system) signal θ_a.

2.2 *Behavioral Modes*

The behavior of a component can vary depending on which *behavioral mode* the component is in. Different types of faults are typically considered to be different behavioral modes. Examples of behavioral modes for a sensor are no-fault, short-cut, bias, and unknown fault. Instead of having these long names for different behavioral modes, it is often practical to use short abbreviations like NF for no-fault, B for bias, UF for unknown fault etc. Further on, we will write $c = UF$ to indicate that component c is in the behavioral mode UF.

For the gas-flow system, the four components are assumed to have the following possible behavioral modes:

Component	**Possible Behavioral Modes**
t	NF
w	NF, SG, UF
p	NF, SG, UF
v	NF, S, SO, SC, UF

where NF means no fault, SG short to ground, UF unknown fault, S stuck, SO stuck open, and SC stuck closed. Note that the tube t is assumed to always be fault free.

As said above, components are described using relations. That is, for each component c there is a set of relations $\{e_{i_c}, e_{i_c+1}, e_{i_c+2}, \ldots\}$ describing the behavior of that component. The validity of each relation can in some cases depend on which behavioral mode the component is in. It can for example be the case that a relation $W_s = W$ holds if component w is in behavioral mode NF, i.e. if $w = NF$, but not necessarily if $w = UF$. This means that together with each relation e_{i_c}, there is an assumption Ass e_{i_c} of the type $c = F_1$ (or a disjunction $c = F_1 \vee c = F_2 \vee \ldots$) that must be fulfilled before the relation e_{i_c} can be assumed to hold, i.e. Ass $e_{i_c} \rightarrow e_{i_c}$. If a relation e_{i_c} is always valid, the assumption Ass e_{i_c} becomes Ass $e_{i_c} = \neg\bot$.

The assumptions and the relations for all components of the gas-flow system are as follows:

	Assumption	**Relation**	
Flow-pipe:			
	$\neg\bot$	$W - f(\theta, P) = 0$	(1)
	$\neg\bot$	0 kg/min $< W <$ 10 kg/min	(2)
	$\neg\bot$	500 kPa $< P <$ 2000 kPa	(3)
Gas-flow sensor:			
	$w = NF$	$W_s = W$	(4)
	$w = SG$	$W_s = 0$	(5)
Pressure sensor:			
	$p = NF$	$P_s = P$	(6)
	$p = SG$	$P_s = 0$	(7)
Valve actuator:			
	$v = NF$	$\theta_a = \theta$	(8)
	$v = S$	$\theta = \tau$	(9)
	$v = SO$	$\theta = 0$ deg	(10)
	$v = SC$	$\theta = 90$ deg	(11)
	$\neg\bot$	0 deg $\leq \theta \leq$ 90 deg	(12)

As seen, the model contains both equations and inequalities. The inequalities are for this system not dependent on any assumptions and are therefore assumed to always hold. The constant τ is unknown and represents the angle of the valve when it is stuck. Noise has not been considered in this example but if desirable, a noise term can be added to each model relation (see Section 6).

2.3 *The System and System Behavioral Modes*

The total model $\mathcal{M}$ of the system to be diagnosed is the union of all relations describing the components. Further on it needs to be specified which of the external variables that are possible to observe for the diagnosis system. The observation vector is denoted z which can be samples from one or several time instances. For the gas-flow system the observed variables are W_s, θ_a, and P_s. If time is considered and the observation vector is collected over a time window of e.g. length 2, $z(t)$ would be $z(t) = [W_s(t-1), \theta_a(t-1), P_s(t-1), W_s(t), \theta_a(t), P_s(t)]$.

As well as defining behavioral modes for the components, we can define *system behavioral-modes*. A system behavioral-mode completely determines the behavioral mode of all components in the system. This is sometimes also called a *mode assignment*. For the gas-flow system, the tube has no alternative behavioral modes and does therefore not need to be included explicitly in a system behavioral-mode. This means that a system behavioral-mode could be for example $w = NF \wedge p = SG \wedge v = UF$, meaning that component w is in behavioral mode NF, p in SG, and v in UF. Other examples of system behavioral-modes are $w = NF \wedge p = NF \wedge v = NF$ and $w = SG \wedge p = NF \wedge v = NF$. An alternative representation is to write a system behavioral-mode using a tuple, e.g. $\langle NF, SG, UF \rangle$.

Like component behavioral-modes, we can use abbreviations to denote system behavioral-modes. This is especially practical when only single-faults are considered. For example for the gas-flow system, the system behavioral modes $w = NF \wedge p = NF \wedge v = NF$ and $w = SG \wedge p = NF \wedge v = NF$ can be written **NF** and $\mathbf{SG}_w$. To say that the system is in for example behavioral mode **NF**, we will write $sys = \mathbf{NF}$.

[1] Another alternative, not exploited here, is to describe connections between components explicitly by using equations, e.g. see the object-oriented modelling-language Modelica.

3. DIAGNOSTIC TESTS

A *diagnosis system* is assumed to consist of a set of *diagnostic tests* which is a special case of a general *statistical hypothesis test* (Casella and Berger, 1990). This idea has in earlier papers been described as *structured hypothesis tests* (Nyberg, 2002). We will in this section discuss diagnostic tests and later, in Section 4, describe how several diagnostic tests are combined to form a diagnosis system.

To define a diagnostic test we need the notion of a *test quantity* $T_k(z)$ which is a function from the observations z to a scalar value. A diagnostic test for a noise-free model can then be defined as follows:

Definition 1. (Diagnostic Test). Let Φ_k be a logical expression in behavioral modes. A *diagnostic test* δ_k for the null hypothesis $H_k^0 : \Phi_k$ is a hypothesis test consisting of a test quantity $T_k(z)$ and a rejection region $\mathcal{R}_k$ such that

$$\Phi_k \rightarrow T_k \in \mathcal{R}_k^C \tag{13}$$

where $\mathcal{R}_k^C$ is the complement of $\mathcal{R}_k$.

This definition will in Section 6 be generalized to the noisy case.

The complement of the null hypothesis is called the *alternative hypothesis* and denoted $H_k^1 : \neg\Phi_k$. Definition 1 means that if $T_k(z) \in \mathcal{R}_k$, Φ_k can not hold. This is the same thing as saying that the null hypothesis H_k^0 is *rejected* and the alternative hypothesis H_k^1 is accepted. The expression Φ_k becomes in this case a so called *conflict* (Kleer and Williams, 1987), i.e. an expression in behavioral modes that is in conflict with the observations.

For the gas-flow example, consider a diagnostic test δ_1 for the null hypothesis $H_1^0 : (p = SG)$, i.e. $\Phi_1 = (p = SG)$. From the model relation (7), we have that $(p = SG) \rightarrow P_s = 0$. This means that a test quantity $T_1(z) = P_s$ and a rejection region $\mathcal{R}_k = [-0.1, 0.1]^C$ implies that $(p = SG) \rightarrow P_s = 0 \rightarrow T_1(z) = P_s = 0 \in \mathcal{R}_k^C$. That is, these choices of $T_1(z)$ and $\mathcal{R}_k$ fulfill the criterion (13) for being a diagnostic test for $H_1^0 : (p = SG)$. When $|T_1(z)| > 0.1$ we reject the null hypothesis $\Phi_1 = (p = SG)$ and draw the conclusion $\neg(p = SG) \simeq (p = NF) \vee (p = UF)$. Note that to evaluate $\simeq$, the assumption that $p \in \{NF, SG, UF\}$ must also be used. From now on when $\simeq$ is used, it will always implicitly be assumed that components must be in exactly one of the behavioral modes.

No conclusion is drawn from a test in which the null hypothesis has not been rejected. That is, to reject null hypotheses is the only way the diagnosis system can draw conclusions. Note that it is usually not true that Φ_k holds when $H_k^0 : \Phi_k$ is not rejected. It would sometimes be possible to assume something else. However, it is in general difficult (or impossible) to construct $T_k(z)$ and $\mathcal{R}_k$ so that such a conclusion can be drawn when the null hypothesis is not rejected.

Another reason why no conclusion is drawn when the null hypothesis is not rejected is that it is not needed. If there is a conclusion that really can be drawn from $T_k \in \mathcal{R}_k^C$, it is always possible to add another diagnostic test to the diagnosis system such that this conclusion can be drawn anyway. For example consider the test δ_1 defined above. When $|T_1(z)| \leq 0.1$ and we do not reject the null hypothesis, it would be tempting to draw the conclusion $\neg(p = NF)$. This is in principle correct because the model also contains the knowledge 500 kPa $< P <$ 2000 kPa so when $T_1(z) = P_s = 0$ it can not hold that $p = NF$. The suggested framework does not allow us to draw a conclusion when a null hypothesis is not rejected, but this desired conclusion can be obtained if we instead add another test δ_2 with $\Phi_2 = (p = NF)$, $T_2(z) = T_1(z)$, and $\mathcal{R}_2 = \mathcal{R}_1^C$.

3.1 *Diagnostic Tests and the Model*

The idea of *model* based diagnosis is to utilize the model $\mathcal{M}$ in the construction of the diagnostic tests. For each diagnostic test δ_k, not necessarily the whole model $\mathcal{M}$ is utilized. Instead only a subset $M_k \subseteq \mathcal{M}$ is considered. The purpose of the diagnostic test δ_k is then to test the validity of the null hypothesis Φ_k, by comparing M_k with the observations. The comparison is done via the test quantity $T_k(z)$ and the rejection region $\mathcal{R}_k$. This means that, in addition to Φ_k, $T_k(z)$, and $\mathcal{R}_k$, also a model M_k must be considered when constructing a diagnostic test. Below we will discuss how the items M_k, Φ_k, $T_k(z)$, and $\mathcal{R}_k$ should be related so that at least the basic requirement (13) is fulfilled. First however, the operator Ass needs to be generalized and a new operator Mod needs to be defined.

In Section 2.2, the notion $\mathrm{Ass}\ e_i$ was used to pick out the assumption for a certain model relation e_i. Here we will use Ass to pick out the assumption also for a set of model relations. If $M_k = \{e_1, e_2, e_3\}$ then $\mathrm{Ass}\ M_k$ means $\mathrm{Ass}\ M_k = \mathrm{Ass}\ e_1 \wedge \mathrm{Ass}\ e_2 \wedge \mathrm{Ass}\ e_3$.

To pick out parts of the model, valid for certain behavioral modes, it is useful to introduce an operator Mod. Given a system behavioral mode ϕ, the expression $\mathrm{Mod}\ \phi$ picks out all relations e_i from the whole model $\mathcal{M}$ such that $\phi \rightarrow \mathrm{Ass}\ e_i$. An example is $\mathrm{Mod}(sys = \mathbf{NF})$ that would pick out relations (1), (2), (3), (4), (6), (8) and (12).

With the Ass and Mod operators we are now able to formulate two guidelines for ensuring that the requirement (13) is fulfilled.

a) The model relations $M_k = \{e_{k1}, e_{k2}, \dots\}$ and the null hypothesis Φ_k should satisfy

$$\Phi_k \rightarrow \mathrm{Ass}\ M_k \tag{14}$$

or even better, $\Phi_k = \mathrm{Ass}\ M_k$.

b) The model relations M_k, the test quantity $T_k(z)$, and the rejection region $\mathcal{R}_k$ should satisfy

$$\big(\exists x : M_k(x, z)\big) \rightarrow T_k(z) \in \mathcal{R}_k^C \tag{15}$$

where the expression $\exists x : M_k(x, z)$ means that the model relations M_k can be fulfilled with a given z.

Note that by definition of $\mathrm{Ass}\ M_k$, it holds that $\mathrm{Ass}\ M_k \rightarrow \exists x : M_k(x, z)$. This means that if the guidelines (a) and (b) are followed, it holds that

$$T_k(z) \in \mathcal{R}_k \rightarrow \neg\exists x : M_k(x, z) \rightarrow \neg\mathrm{Ass}\ M_k \rightarrow \neg\Phi_k \tag{16}$$

That is, when the test quantity is within the rejection region, we can draw the conclusion that Φ_k can not hold. This expression is equivalent to the requirement (13) so the design goal has been achieved. Note that if $\Phi_k = \mathrm{Ass}\ M_k$ instead of only (14), a stronger conclusion can in general be drawn in (16). As said above, $\Phi_k = \mathrm{Ass}\ M_k$ is therefore normally a better choice than (14).

For an example, consider the gas-flow system and a test δ_3. For the choices $\Phi_3 = (w = NF)$, $T_3(z) = W_s$, $\mathcal{R}_3^C = [-2, 12]$, and $M_3 = \{0 < W < 10, W_s = W\}$, it holds that $\Phi_3 = (w = NF) = \mathrm{Ass}\ M_3$ and that

$$\exists W : M_3(W, W_s) \leftrightarrow 0 < W_s < 10 \rightarrow$$
$$\rightarrow W_s \in [-2, 12] \leftrightarrow T_3(z) \in \mathcal{R}_3^C$$

Guidelines (a) and (b) are fulfilled and therefore also the requirement (13).

4. THE DIAGNOSIS SYSTEM

A diagnosis system using the principle of consistency based diagnosis takes the observations and tries to conclude which behavioral modes that can explain the observations. Let the output from a diagnosis system be called *diagnostic statement* and denoted $\mathcal{S}$.

Formally a diagnosis system based on structured hypothesis tests can be defined as follows:

Definition 2. (Diagnosis System). A *diagnosis system* is a set of diagnostic tests, i.e. $\{\delta_1, \delta_2, \dots\}$, together with the procedure to form the diagnostic statement $\mathcal{S}$ defined as

$$S = \bigwedge_{H_k^0 \text{ rejected}} \neg\Phi_k \tag{17}$$

4.1 *Strategies for Designing Diagnosis Systems*

To design a diagnosis system consists of finding the set of diagnostic tests to be included, and also for each diagnostic test, a test quantity $T_k(z)$, a rejection region $\mathcal{R}_k$, and a null hypothesis Φ_k. We will here study two different strategies for finding these items. The first starts from a given set of null hypotheses Φ_k, and the second from the model $\mathcal{M}$ of the system to be diagnosed.

4.2 *Starting From a Given Set of Null Hypotheses*

One way of starting the design of a diagnosis system is simply to decide which null hypotheses to test, and then construct a suitable test quantity and rejection region for each hypothesis test. One straightforward strategy is for example to have one diagnostic test for each of the system behavioral-modes. This is especially attractive when only single faults are considered. For example, if the possible system behavioral-modes are **NF**, **F1**, **F2**, and **F3**, the four null hypotheses become

$$\begin{aligned} H_{k_1}^0 &: \ sys = \mathbf{NF} \\ H_{k_2}^0 &: \ sys = \mathbf{F1} \\ H_{k_3}^0 &: \ sys = \mathbf{F2} \\ H_{k_4}^0 &: \ sys = \mathbf{F3} \end{aligned}$$

To fulfill (13) it is suggested to follow the guidelines (a) and (b) above. The guidelines will then tell us how to choose M_k, namely any set such that (14) is fulfilled. The test quantity $T_k(z)$ and the rejection region $\mathcal{R}_k$ should then be selected to fulfill (b).

For example consider again the gas flow example and assume that we want to design a diagnostic test for the null hypothesis $H_4^0 : sys = \mathbf{NF}$. Now select the set M_4 to consist of relations (1), (4), (6), and (8). Then $\text{Ass}\, M_4 = (w = NF) \wedge (p = NF) \wedge (v = NF)$ which means that $\Phi_4 = (sys = \mathbf{NF}) = (w = NF) \wedge (p = NF) \wedge (v = NF) = \text{Ass}\, M_4$, and formula (14) is therefore trivially fulfilled. By eliminating the unknown variables W, θ, and P, the four relations are reduced to $W_s - f(\theta_a, P_s) = 0$. By then selecting $T_4(z) = W_s - f(\theta_a, P_s)$ and $\mathcal{R}_k = [-0.1, 0.1]^C$, it is ensured that formula (15) is fulfilled.

4.3 *Starting From the Model*

The idea of this strategy is to start out from the model relations and investigate which relations that can be grouped together to form models possible to test in diagnostic tests. That is, we have to find those subsets M_k that are meaningful to check for validity. The null hypothesis $H_k^0 : \ \Phi_k$ will then be chosen as $\Phi_k = \text{Ass}\, M_k$. In this way the relation (14) will of course be fulfilled. Then the selection of the test quantity $T_k(z)$ and the rejection region $\mathcal{R}_k$ should follow (b).

One requirement on the subset M_k is that there must be some z such that ideally (i.e. without noise), there exists some observation z such that the relations M_k cannot all be fulfilled, i.e. $\exists z \forall x : \neg M_k(x, z)$. If this requirement is not fulfilled, the test quantity would always be zero, or close to zero, and the test would make no sense. Another requirement is that $\text{Ass}\, M_k \not\equiv \perp$. If this requirement would not be fulfilled it would hold that $\neg\Phi_k \equiv \neg\perp$. This means that the result of rejecting a null hypothesis would be that we can draw the conclusion $\neg\perp$, i.e. the test can never provide any information.

The question that remains is how to find the subsets M_k such that these two requirements are satisfied. Given some natural assumptions about the model, the problem of finding suitable subsets M_k can often be solved by only studying the structural properties of the model. This is not the topic of this paper but the interested reader is referred to (Krysander and Nyberg, 2002).

Now consider the gas-flow system and assume that the following subsets M_k with their corresponding assumptions $\text{Ass}\, M_k$ have been found possible to validate:

Relations M_k	**Assumption** Ass M_k
(7)	$p = SG$
(3), (6)	$p = NF$
(2), (4)	$w = NF$
(1), (4), (6), (8)	$w = NF \wedge p = NF \wedge v = NF$
(5)	$w = SG$
(1), (4), (6), (9)	$w = NF \wedge p = NF \wedge v = S$
(1), (4), (6), (10)	$w = NF \wedge p = NF \wedge v = SO$
(1), (4), (6), (11)	$w = NF \wedge p = NF \wedge v = SC$

As said above, Φ_k is then chosen as $\Phi_k = \text{Ass}\, M_k$. By eliminating unknown variables, each model M_k can be written with one, so called *consistency relation*, containing only observations. These consistency relations are:

Relations M_k	**Consistency Relation**
(7)	$P_s = 0$
(3), (6)	500 kPa $< P_s <$ 2000 kPa
(2), (4)	0 kg/min $< W_s <$ 10 kg/min
(1), (4), (6), (8)	$W_s - f(\theta_a, P_s) = 0$
(5)	$W_s = 0$
(1), (4), (6), (9)	$W_s - f(\tau, P_s) = 0$
(1), (4), (6), (10)	$W_s - f(0, P_s) = 0$
(1), (4), (6), (11)	$W_s = 0$

With these consistency relations, test quantities $T_k(z)$ and rejection regions $\mathcal{R}_k$ can easily be constructed to fulfill (b).

5. CONNECTION TO FDI METHODS

FDI methods presented in the literature, have focused mostly on residual generation and how disturbances and faults are to be decoupled. A residual is a signal that is zero in the fault-free case, and to use residuals is the most common way to construct test quantities within the field of FDI. The reason to decouple disturbances is to avoid false alarms, and the reason to decouple faults is to obtain residuals that are sensitive to different subsets of faults, so that isolation can be performed. From a residual r_k, a test quantity can for example be formed as $T_k = |r_k|$ or $T_k = \sum_{t=t_0}^{t_0+N} r_k^2(t)$.

Consider a linear system, typically found in FDI literature:

$$\dot{x} = \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix} x + \begin{bmatrix}0\\ 2\end{bmatrix} u_a + \begin{bmatrix}1\\ 1\end{bmatrix} d + \begin{bmatrix}2\\ 1\end{bmatrix} f_1 + \begin{bmatrix}1\\ 0\end{bmatrix} f_2 \quad (18a)$$

$$y = \begin{bmatrix}1 & 0\\ 0 & 1\\ 1 & 1\end{bmatrix} x + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix} d \quad (18b)$$

where x is the dynamic state, u_a the actuator signal, y the sensor signals, and d an unknown disturbance signal. The signals f_1 and f_2 are used to model two different faults of the system and are non-zero only if the corresponding fault is present. The system itself, denoted c, is considered to have three possible behavior modes: NF, $F1$, and $F2$. As seen, no actuator or sensor faults have been considered. The model $\mathcal{M}$ for this system, rewritten using the modelling framework suggested here, becomes

Assumption	**Relation**	
System:		
$c = NF$	$\dot{x}_1 = x_1 + x_2 + d$	(19)
$c = NF \vee c = F1$	$\dot{x}_1 = x_1 + x_2 + d + 2f_1$	(20)
$c = NF \vee c = F2$	$\dot{x}_1 = x_1 + x_2 + d + f_2$	(21)
$c = NF \vee c = F2$	$\dot{x}_2 = x_1 + 2u + d$	(22)
$c = NF \vee c = F1$	$\dot{x}_2 = x_1 + 2u + d + f_1$	(23)
Actuator:		
$\neg\bot$	$u = u_a$	(24)
Sensor 1:		
$\neg\bot$	$y_1 = x_1$	(25)
Sensor 2:		
$\neg\bot$	$y_2 = x_2$	(26)
Sensor 3:		
$\neg\bot$	$y_3 = x_1 + x_2 + d$	(27)

The goal now is to find some residual for the system (18). In all residuals, the unknown disturbance d must be decoupled. To facilitate isolation, the goal is also to decouple different faults in different residuals. By linear-algebra manipulations of the system (18) (e.g. see (Frisk and Nyberg, 2001)), a number of residual generators can be found (here in the form of so called *parity relations*), for example:

$$\begin{aligned} r_1 &= - \dot{y}_1 + y_3 \\ r_2 &= - 4u - \dot{y}_1 - 2y_2 + 2\dot{y}_2 - y_3 \\ r_3 &= 2u + y_1 - \dot{y}_2 - y_2 - y_3 \end{aligned}$$

By carefully studying the formula of each residual, it can be realized that the sensitivity to the faults is according to the second column of the following table:

	NF	$F1$	$F2$	M_k	Ass M_k
r_1	0	X	X	(19),(24-27)	$c = NF$
r_2	0	0	X	(20),(23),(24-27)	$c{=}NF \vee c{=}F1$
r_3	0	X	0	(22),(24-27)	$c{=}NF \vee c{=}F2$

A “0” means that when the behavioral mode of the column is present the residual of that row will be zero. “X” means that the residual will be zero or non-zero. That is, in residual r_2, the fault signal f_1 has been decoupled, and in r_3, f_2 has been decoupled.

To see the relationship with the framework presented here, we have to investigate exactly which equations that have been used to form each residual. It turns out that to form residual r_1, i.e. to derive the equation $-\dot{y}_1 + y_3 = 0$, from the equations in the model $\mathcal{M}$, exactly the equations (19), (24), (25), (26), and (27) have to be used. The equations M_k used to derive r_1, r_2, and r_3 can be seen in the third column of the table. The assumptions for each equation set M_k, i.e. Ass M_k, can be seen in the fourth column of the table.

In conclusion, the FDI methods for residual generation, which can be based on e.g. parity relations or observers, can be fully utilized in the framework presented here. By keeping track of exactly which set M_k of equations that are used in the construction of each residual, the expression Ass M_k can be obtained. This is then the only thing that is needed to facilitate isolation in the way proposed in this paper.

6. NOISY SYSTEMS

The relation (13) can normally only hold strictly when the diagnostic test is used together with a noise-free system. If noise is present, (13) has to be replaced by specifying the probability that (13) holds. In statistical hypothesis-testing theory, this is usually written as

$$P(T_k \in \mathcal{R}_k \mid \Phi_k) \leq \alpha \quad (28)$$

That is, the probability of rejecting the null hypothesis Φ_k given that Φ_k holds must be less or equal to a *significance level* α. The idea behind hypothesis testing is to have a significance level that is very small, in fact so small that it is realistic to assume that the formula (13) holds.

In noisy (stochastic) systems, the model $\mathcal{M}$ is only approximate, or alternatively, is exact but includes stochastic terms. Regardless of the view chosen, we can assume that each model equation includes a stochastic term with some probability distribution. For example if the model equation (1) is only approximate, we can indicate this by writing

$$W - f(\theta, P) + n = 0 \quad (29)$$

where n is a noise term with for example Gaussian distribution, i.e. $n \sim N(0, \sigma)$.

Next we study how the guidelines (a) and (b) in Section 3.1 are transformed to the noisy case. First, the guideline (a) is not changed; the only difference is that the model is approximate and not exact. For the guideline (b), equation (15) is replaced by

$$P\big(T_k(z) {\in} \mathcal{R}_k^C \,|\, \exists x, n {:} (M_k(x, z, n), n {\sim} N(0, \Sigma))\big) \geq 1 {-} \alpha' \quad (30)$$

where n is the noise, $N(0, \Sigma)$ the probability distribution of n, and α' a small number. The interpretation is that when the observations are compatible with the model $M_k(x, z, n)$, which is now stochastic, the probability of not rejecting the null hypothesis is very high. Note that the probability distribution of $T_k(z)$ can be obtained from $N(0, \Sigma)$ by propagating the noise n through the model $M_k(x, z, n)$ and $T_k(z)$. Note also that the formula (15) is obtained if $n = 0$ and α' is chosen as $\alpha' = 0$.

For the gas-flow system, assume that equation (1) contains a noise term according to (29), but (4), (6), and (8) do not contain any noise terms. This means that the model M_4 can be reduced to $W_s - f(\theta_a, P_s) + n = 0$. By then selecting $T_4(z) = W_s - f(\theta_a, P_s)$ and $\mathcal{R}_k = [-0.1, 0.1]^C$, the formula (30) becomes

$$P\big(|n| \leq 0.1 \mid W_s {-} f(\theta_a, P_s) {+} n {=} 0, n {\sim} N(0, \sigma)\big) {\geq} 1 {-} \alpha'$$

If we assume that the standard deviation σ is small, then the formula will hold with a small α'. That is, the stochastic version of guideline (b) is fulfilled.

From the definition of Ass M_k, it holds that Ass $M_k \rightarrow \exists x, n : (M_k(x, z, n), n \sim N(0, \Sigma))$. Then by using the formula (14) and (30), we obtain

$$P(T_k(z) \in \mathcal{R}_k^C \mid \Phi_k) \geq 1 - \alpha'$$

which is equivalent to (28), and the α' chosen becomes the significance level. Thus, the stochastic version of the guidelines (a) and (b) will produce diagnostic tests in accordance with the formula (28), which is the stochastic version of the diagnostic test. With (28) fulfilled, it is then realistic to assume that (13) holds which means that we are back to the no-noise case. That is, the framework and principles discussed in this paper for the no-noise case are applicable also for the noisy case.

7. THE DIAGNOSTIC STATEMENT

The goal is that the output from a diagnosis system, i.e. the diagnostic statement, should tell which system behavioral-modes that can explain the given observation z. Such a system behavioral-mode is called a *diagnosis*. Note that this definition of diagnosis is equivalent to the one used in consistency based diagnosis (Hamscher *et al.*, 1992). Ideally the diagnostic statement (17) should say everything about which diagnoses that hold and which that do not hold. However depending on how the diagnosis system is constructed it is not sure that there is this exact relationship between the diagnostic statement and the diagnoses. Therefore we define *candidate* $\mathcal{C}$ as a system behavioral mode such that $\mathcal{C} \to \mathcal{S}$. That is, a candidate is a system behavioral mode that the diagnosis system claims to be able to explain the observations.

For an example, consider the example diagnosis-system constructed in Section 4.3. Assume that the single fault $v = S$ is present, i.e. the valve is stuck in angle τ but the other components are fault-free. Assume that we have an exciting input signal which means that $\theta_a \neq \tau$. If f is assumed monotonic, this implies that the null hypotheses Φ_1, Φ_4, Φ_5, Φ_7, and Φ_8 would be rejected. Using (17), the diagnostic statement $\mathcal{S}$ would be

$$\mathcal{S} = \neg\Phi_1 \wedge \neg\Phi_4 \wedge \neg\Phi_5 \wedge \neg\Phi_7 \wedge \neg\Phi_8 =$$

$$\neg(p = SG) \wedge \neg(w = NF \wedge p = NF \wedge v = NF) \wedge \neg(w = SG)\wedge$$

$$\neg(p = SG) \wedge \neg(w = NF \wedge p = NF \wedge v = SO)\wedge$$

$$\neg(w = NF \wedge p = NF \wedge v = SC)$$

This expression is not so easy to interpret. However, let us transform the expression to *full disjunctive normal form*, which here means a disjunction of conjunctions where each conjunction contains exactly one assignment for each component:

$$\begin{aligned}\mathcal{S} \simeq\ & (p = NF \wedge v = NF \wedge w = UF)\vee \\ & (p = NF \wedge v = SC \wedge w = UF)\vee \\ & (p = NF \wedge v = SO \wedge w = UF)\vee \\ & \vdots \\ & (p = UF \wedge v = UF \wedge w = UF) \qquad (31)\end{aligned}$$

The full expression consists of 17 conjunctions, and each conjunction is a candidate. That is, the full disjunctive normal form is simply a complete list of all candidates produced by the diagnosis system. For a repair technician, this information is usually not precise enough. For efficient repair, more focused information is preferred.

One way of filtering the diagnostic statement is to only consider the "simplest" diagnoses. Formally a *preference* relation can be used as described in (Dressler *et al.*, 1993). If for component c, mode F_c' is preferred ("simpler than") over F_c, we write this $F_c' < F_c$. The relation $\leq$ is then a partial order on the component behavioral modes. The semantics of the preference relation can be for example probability, i.e. $F_c' < F_c$ means that F_c' is more probable than F_c. Another choice is to compare the solution sets of the external variables for each component.

In the gas-flow example we assume the following relations between the behavioral modes:

w	$NF < SG < UF$
p	$NF < SG < UF$
v	$NF < S < UF,\ NF < SO < UF,\ NF < SC < UF$

By applying this preference relation to the diagnostic statement (31), we obtain the following three minimal candidates

$$p = NF \wedge v = NF \wedge w = UF$$

$$p = NF \wedge v = S \wedge w = NF$$

$$p = UF \wedge v = NF \wedge w = NF$$

8. CONCLUSIONS

A new framework for model based diagnosis has been presented. The isolation mechanism follows ideas from AI, namely to include in the model, how the validity of model equations depend on the presence of faults in each component. Isolation is then performed by propagating this information through the diagnosis system by the use of standard logic. This isolation strategy is far more competent than the isolation strategy *structured residuals* (Gertler and Singer, 1990) that is typically used in FDI literature. For example, no special care for isolation of multiple faults is needed.

In contrast to AI, the diagnostic tests are computed off-line as in FDI. It has been shown in Section 5 how standard FDI methods, such as residuals based on parity relations or observers, can be used within the framework. In that case, the powerful isolation mechanism can be fully utilized.

Since the diagnostic tests used are really standard hypothesis tests from statistical hypothesis testing theory, it is guaranteed that noise get a sound treatment. That is, even in a noisy system, faults are correctly isolated. This is not true in for example the method *structured residuals*, see (Nyberg, 1999).

In summary, the framework presented can efficiently handle: fault models, several different fault types (e.g. parameter- and additive faults), more than two behavioral modes per component, general differential-algebraic models, noise, uncertainties, decoupling of disturbances, static and dynamic systems, and isolation of multiple faults.

9. REFERENCES

Casella, G. and R.L. Berger (1990). *Statistical Inference*. Duxbury Press.

Dressler, O., C. Böttcher, M. Montag and A. Brinkop (1993). Qualitative and quantitative models in a model-based diagnosis system for ballast tank systems. In: *Proceedings Int. Conf. on Fault Diagnosis (TOOLDIAG)*. Toulouse, France. pp. 397–405.

Frisk, E. and M. Nyberg (2001). A minimal polynomial basis solution to residual generation for fault diagnosis in linear systems. *Automatica* **37**, 1417–1424.

Gertler, J. and D. Singer (1990). A new structural framework for parity equation-based failure detecation and isolation. *Automatica* **26**(2), 381–388.

Hamscher, W., L. Console and J. de Kleer (1992). *Readings in MODEL-BASED DIAGNOSIS*. Morgan Kaufmann Publishers.

Kleer, J. De and B.C. Williams (1987). Diagnosing multiple faults. *Artificial Intelligence* **32**(1), 97–130.

Krysander, M. and M. Nyberg (2002). Structural analysis utilizing mss sets with application to a paper plant. In: *Thirteenth International Workshop on Principles of Diagnosis*. Semmering, Austria.

Nyberg, M. (2002). Model-based diagnosis of an automotive engine using several types of fault models. *IEEE Transactions on Control Systems Technology* **10**(5), 679–689.

Nyberg, Mattias (1999). Model Based Fault Diagnosis: Methods, Theory, and Automotive Engine Applications. PhD thesis. Linköping University.