Statistical Properties and Design Criterions for AI-Based Fault Isolation

Mattias Nyberg and Mattias Krysander

Department of Electrical Engineering, Linköping University,
SE-581 83 Linköping, Sweden,
{*matny,matkr*}*@isy.liu.se.*

Abstract: Fault diagnosis in the presence of noise and model errors is of fundamental importance. In the paper, the meaning of fault isolation performance is formalized by using the established notion of coverage and false coverage from the field of statistics. Then formal relations describing the relationship between fault isolation performance and the residual related design parameters are derived. For small faults, the measures coverage and false coverage are not applicable so therefore, a different performance criteria, called sub-coverage, is proposed. The performance of different AI-based fault isolation schemes is evaluated and it is shown that the well known principle of minimal cardinality diagnosis gives a very bad performance. Finally, some general design guidelines that guarantee and maximize the fault isolation performance are proposed.

Keywords: fault diagnosis, AI-methods, fault isolation, FDI-methods, noise, model errors, statistics

1. INTRODUCTION

The FDI (Fault Detection and Isolation) problem, as often described within the control community, is to detect and isolate any possible faults given sensor and actuator signals only. A typical solution, see [4, 12, 1], is to use a set of thresholded residuals together with a fault isolation scheme, which, based on the fact that the thresholded residuals respond differently to different faults, isolates the fault.

In a real application, there are typically model errors and noise. This fact limits our ability to construct a diagnosis system that perfectly detects and isolates the present fault. However, there is also design freedom available such as the threshold levels, the set of residuals to be included, and which isolation strategy to use. Thus, under the premises of noise and model errors, the design freedom should be utilized such that the ability of detecting and isolating faults is optimized.

The discussion above reveals first of all, that there is a need for an exact measure of FDI performance. Secondly it is important to understand how this FDI performance changes when different design parameters are changed. In the literature, only a few studies have addressed these issues. In [9], FDI performance was studied in the framework of structured hypothesis tests. In [3] these issues were posed as open questions.

In several works [10, 13, 3], it has been recognized that fault isolation in FDI can be solved by using algorithms developed within the field of AI, see [7, 14]. Advantages of these AI algorithms, compared to their counterpart from the control community, e.g. [4], are that they can easily handle multiple faults and their computational efficiency. Because of these advantages we have in the present paper chosen to focus entirely on fault isolation algorithms from AI. However, the results can be easily generalized to cover fault isolation techniques from the control community such as *structured residuals* [4].

In the paper, a first contribution is to formalize what we mean by FDI performance, especially for noisy and uncertain systems. For this we use the established notion of *coverage* and *false coverage* from the field of statistics. Then as a second contribution, we derive formal relations describing the relationship between FDI performance and the residual related design parameters. Further it is noted that a different performance criteria is needed for small faults, and we therefore introduce a third performance measure called *sub-coverage*. We then discuss the intrinsic FDI performance of different AI-based fault isolation schemes. It is notable that the well known principle of *minimal cardinality diagnosis* gives a very bad performance for the case of small faults. Based on the performance measure and investigations, we develop some general design guidelines that, if followed, guarantee and maximize the fault isolation performance. Finally we illustrate the theory and the guidelines on a small application example.

2. STOCHASTIC VIEW ON DIAGNOSIS

In many papers, both from the control community [4, 12, 1] and especially in AI [6, 3], the systems to be diagnosed are assumed not to contain noise. This means that an observation in the model is either deterministic given the states, or completely unknown, depending on if a fault is present and also which fault that is present. The view taken here is that a system contains stochastic parts which implies that, given the states, observations have probability distributions rather than exact values. Based on this idea we will below give a basic stochastic framework for diagnosis.

2.1 The System

The system to be diagnosed consists of a number of components, and we assume here that the behavioral mode of a component is either non-faulty or faulty, abbreviated NF and F respectively. The behavioral mode of the complete system, called *system behavioral mode* or simply *mode*, can be described by a vector of length equal to the number of components, e.g. in a system with 5 components the system behavioral mode could be $[NF, F, NF, NF, F]$.

Further, we assume that the system has a vector-valued trajectory z which is possible to observe. The vector z includes measured sensor values and actuated control values. Since we have a stochastic view on diagnosis, we consider z to be a random variable. For each system behavioral mode, we assume that z has exactly one given pdf (probability density function). Later in Section 6 we will relax this assumption.

2.2 The Diagnosis System

We consider a *diagnosis system* to be a system that takes an *observation* as input and computes *candidates*, i.e. a set C of system behavioral modes, as output. The candidate set C is assumed to be a function of the observation and supposed to be the system behavioral modes that are likely explanations of the observation.

Formally we define observation as follows.

Definition 1. (Observation). An *observation* $z_{\mathcal{T}}$ of z is samples of z at times specified by the index set $\mathcal{T}$.

Here we allow $\mathcal{T}$ to be a finite or infinite set. Examples of $\mathcal{T}$ are $\mathcal{T} = \{0\}$, $\mathcal{T} = \{0, 1, 3\}$, $\mathcal{T} = [0, 2]$, and $\mathcal{T} =]\infty, 0]$.

Since z is a random variable, and $z_{\mathcal{T}}$ is a function of z, also $z_{\mathcal{T}}$ will be a random variable. Since the pdf of z was assumed to be given uniquely for each system behavioral mode b, also $z_{\mathcal{T}}$ will have a unique pdf denoted $f_b(z_{\mathcal{T}})$. Lastly, since the candidate set C is a function of the observation $z_{\mathcal{T}}$, also C is a random variable which for each mode will have a unique pdf.

3. STATISTICAL PERFORMANCE MEASURES OF DIAGNOSIS SYSTEMS

Two performance measures of set estimators known from statistical decision making theory [2] will here be introduced as performance measures for diagnosis systems regarding their fault isolation capability. Note that in these performance measures, fault detection becomes a special case of fault isolation so we will refer only to fault isolation performance from now on.

3.1 Coverage Probability

Suppose that we want to diagnose a system that is operating in an unknown mode. It is almost never possible for a diagnosis system to exactly determine the present mode. A more realistic objective is that the candidate set C should at least with some high probability contain the present mode and the first performance measure formalizes this idea.

Definition 2. (Coverage Probability). Given a diagnosis system computing the candidate set C, the *coverage probability* is a function of b given by

$$P(b \in C \,|b) \tag{1}$$

Practical Relevance of Coverage

Let $\mathbf{NF}$ denote the fault free system behavioral mode. False alarm can formally be described as the negation of coverage with respect to the mode $\mathbf{NF}$. False alarms lead to expensive and unnecessary troubleshooting. Further, they degrade both the perceived product quality and the confidence in the diagnosis system. Therefore false alarms are in general not accepted in industrial applications.

Consider next the event $b \notin C$ in the case that the present mode is b where $b \neq \mathbf{NF}$. If the user of the diagnosis result takes action based on the fact that b can not be the present mode, severe and expensive mistakes might be done. For example, if a repair technician excludes the possibility that b is the present mode, he will replace non-faulty parts and still not succeed with his repair mission.

From this discussion it is clear that lack of coverage is in general not acceptable in industrial applications.

3.2 False Coverage Probability

It is not sufficient to evaluate the isolation performance of a diagnosis system by using only its coverage probabilities. Ideally we also want to exclude all modes that are not the present mode.

Definition 3. (False Coverage Probability). Given a diagnosis system computing the candidate set C, the *false coverage probability* is a function of b and b' given by

$$P(b' \in C \,|b), \quad \text{where } b' \neq b \tag{2}$$

Note that, in contrast to coverage probability which is a function defined on each mode, the false coverage probability is a function defined on each non-equal pair of modes.

Practical Relevance of False Coverage

False coverage means that $b' \in C$ even though another mode b is the present one. This is of course not a desired situation since it implies that the user of the diagnosis result has to undertake unnecessary safety or repair actions or to convey further analysis to exclude the mode b'. However we consider it not as serious as lack of coverage.

4. DIAGNOSIS SYSTEMS USING AI-BASED FAULT ISOLATION

As said in the introduction, we consider diagnosis systems consisting of a set of diagnostic tests together with a fault isolation scheme using techniques from the field of AI. Further, we consider diagnostic tests in the view of hypothesis testing in accordance with [9]. It should be noted that this view is compatible with traditional fault isolation techniques from both FDI and AI, see [3].

The main idea is the following. Each diagnostic test δ_k is a hypothesis test with a null hypothesis H_0^k and a rejection region R^k. The diagnostic test takes an observation $z_{\mathcal{T}}$ as input and generates a binary decision as output as follows. If $z_{\mathcal{T}} \in R^k$, then H_0^k is rejected, otherwise H_0^k is not rejected. The null hypothesis H_0^k is here represented as a set of system behavioral modes. When the null hypothesis is rejected, the conclusion from the diagnostic test is that none of the modes in H_0^k is the one that has generated the observation $z_{\mathcal{T}}$, i.e. the present mode must be in the complement set ${H_0^k}^C$. Using AI terminology, a rejected null hypothesis H_0^k is a so called conflict.

In the isolation scheme, the conclusions from the individual diagnostic tests are merged. In its simplest form, the isolation scheme is a simple intersection of the conclusions from the tests, i.e.

$$C = \bigcap_{\substack{k \\ H_0^k \text{ is rejected}}} {H_0^k}^C \tag{3}$$

This principle has been used in both FDI and AI [9, 3] even though more efficient representations and computations have been utilized.

For an example, let $\mathbf{F2}$ denote the system behavioral mode with a fault in component 2 only, let $\mathbf{F12}$ denote the system behavioral mode with faults in components 1 and 2 only, etc. Then consider the following table which we call *decision structure*:

	NF	**F1**	**F2**	**F3**	**F12**	**F23**	**F13**	**F123**
δ_1	0	X	X	0	X	X	X	X
δ_2	0	X	0	0	X	X	X	X
δ_3	0	0	X	X	X	X	X	X

(4)

A 0 in row i and column j means that the mode of column j is a member of the null hypothesis of the test corresponding to row i, i.e. H_0^i. Assume that $\mathbf{F2}$ is the present mode and that the

null hypotheses of the tests δ_1 and δ_3 have been rejected. Then, according to (3),

$$\begin{aligned} C &= H_0^{1C} \cap H_0^{3C} = \\ &= \{\mathbf{F1}, \mathbf{F2}, \mathbf{F12}, \mathbf{F23}, \mathbf{F13}, \mathbf{F123}\} \cap \\ &\qquad \{\mathbf{F2}, \mathbf{F3}, \mathbf{F12}, \mathbf{F23}, \mathbf{F13}, \mathbf{F123}\} = \\ &= \{\mathbf{F2}, \mathbf{F12}, \mathbf{F23}, \mathbf{F13}, \mathbf{F123}\} \end{aligned} \quad (5)$$

A problem with the fault isolation scheme (3), and as seen even in this small example, is that the candidate set C will in general be very large and include many other modes in addition to the present one. This problem is well known and has in the field of AI been solved by, in a second step [1] , filtering out less likely modes from C. This is often called *focusing* and is based on the idea of a preference relation $\leq_p$ defined on the set of system behavioral modes.

For example, in (5), if single faults are preferred compared to multiple faults, the result is a focused set of candidates $C_F = \{\mathbf{F2}\}$, which is actually the perfect result since $\mathbf{F2}$ was the mode assumed to be present. Formally, the set C_F can be defined as

$$C_F = \{b \in C \mid \neg \exists b' \in C : b' >_p b\} \quad (6)$$

The preference relation $\leq_p$ can be defined using different principles of which the concepts of *minimal diagnoses* [7, 14, 5] and *minimal cardinality diagnoses* [15] are the two most common. In Section 7, these preference relations and also the case without focusing, i.e. (3), will be compared with respect to the fault isolation performance measures presented in Section 3.

5. BOUNDS FOR THE PERFORMANCE MEASURES

In this section we will present bounds for the performance measures presented in Section 3. The idea of these bounds is to estimate the performance measures (1) and (2) by using only the performance of the individual diagnostic tests. The performance of each diagnostic test is specified in terms of the probability $P(\text{reject } H_0^k \mid b)$ which, in the field of statistics, is called *power function* [2]. For convenience we will use the shorter writing $P(\text{rej}_k \mid b)$.

The rationale behind bounds of this type is that the design freedom in designing diagnosis systems of the type described in Section 4 lies in the selection and construction of the diagnostic tests. Thus, it is critical to know the relationship between the performance of the individual tests and the performance of the complete diagnosis system. By utilizing these bounds, performance requirements on the individual tests can be derived from diagnosis-system performance requirements.

In the bounds we will use the notation Ω_b for the index set of tests which contain mode b in its null hypothesis, i.e.

$$\Omega_b = \{i | b \in H_0^i\} \quad (7)$$

In the decision structure, Ω_b is the rows with 0 in column b. For example, in (4), $\Omega_{\mathbf{F3}} = \{1, 2\}$.

Basic probability theory gives the general relations $P(A) + P(B) - 1 \leq P(A \wedge B) \leq \min(P(A), P(B))$ and $\max(P(A), P(B)) \leq P(A \vee B) \leq P(A) + P(B)$ for two arbitrary events A and B. Using these relations we can derive the bounds given in the following theorem.

Theorem 1. Let B be the set of modes that are more preferred than mode b, i.e. $B = \{\bar{b} | \bar{b} >_p b\}$. If $\Omega_{\bar{b}} \subseteq \Omega_b$ for some $\bar{b} \in B$, then

$$P(b \in C_F \mid b') = 0 \quad (8)$$

for all b'. Otherwise, for mode b' it holds that

$$\begin{aligned} 1 - |B| - \sum_{k \in \Omega_b} P(\text{rej}_k | b') + \sum_{\bar{b} \in B} \max_{j \in \Omega_{\bar{b}} \setminus \Omega_b} P(\text{rej}_j | b') \\ \leq P(b \in C_F \mid b') \leq \\ \min \Big(1 - \max_{k \in \Omega_b} P(\text{rej}_k | b'), \min_{\bar{b} \in B} \sum_{j \in \Omega_{\bar{b}} \setminus \Omega_b} P(\text{rej}_j | b')\Big) \end{aligned} \quad (9)$$

The proofs of the results in this paper can be found in [11]. Note that no assumption about the correlation between the response of different tests has been made in the theorem above.

From Theorem 1 a number of bounds can be derived both for coverage probability and false coverage probability. For example if a bound for coverage probability in the case of no focusing is needed, let $b = \bar{b}$ and $B = \emptyset$.

Later in the paper we will use the following simplified upper bound for false coverage probability.

Corollary 1. (False Coverage Probability). It holds that

$$P(b \in C_F \mid b') \leq 1 - \max_{k \in \Omega_b} P(\text{rej}_k | b') \quad (10)$$

Next, by using the assumption

$$P(\text{rej}_k | b) = 0, \text{ for all } b \in H_0^k \quad (11)$$

a simplified lower bound for coverage probability can be derived. Note that (11) implies that we assume that the false alarm probability is zero.

Corollary 2. (Coverage Probability). Assume that (11) holds and let B be defined as in Theorem 1. If $\Omega_{\bar{b}} \subseteq \Omega_b$ for some $\bar{b} \in B$, then

$$P(b \in C_F \mid b) = 0 \quad (12)$$

for all b. Otherwise, it holds that

$$1 - |B| + \sum_{\bar{b} \in B} \max_{j \in \Omega_{\bar{b}} \setminus \Omega_b} P(\text{rej}_j | b) \leq P(b \in C_F \mid b) \quad (13)$$

6. RELAXING THE ASSUMPTION OF UNIQUE DISTRIBUTIONS

In Section 2.1 we assumed that $z_{\mathcal{T}}$, and consequently C and C_F, have exactly one given pdf for each mode b. This assumption is quite restrictive since it requires that the behavior of a fault is relatively well known. Thus it is desirable to relax this assumption. We do this here by assuming that for a specific mode b, the random variable $z_{\mathcal{T}}$ has a pdf in a set Φ_b.

The next issue is the performance measures presented in Section 3. For example, the coverage probability $P(b \in C_F | b)$ is no longer well defined since the fact that b is the true mode does not give a single distribution for $z_{\mathcal{T}}$ and consequently not for C_F. Our solution to this problem is to instead consider a coverage probability conditioned on one specific distribution in the set Φ_b. Thus we write

$$P(b \in C_F | z_{\mathcal{T}} \sim f_b(z_{\mathcal{T}})) \qquad f_b(z_{\mathcal{T}}) \in \Phi_b \quad (14)$$

For convenience we will mostly write $P(b \in C_F | f_b(z_{\mathcal{T}}))$ instead of (14). When using the coverage probability measure (14), and only the set Φ_b is specified, we do not get a single coverage probability for a specific mode b but instead a set, possibly infinite, of coverage probabilities. Thus, the next question is how to use such a performance measure.

First, note that a mode b may contain both small and large faults. For example consider the mode bias of a sensor. There are both small biases, close to zero and large ones. Because we consider stochastic noisy systems, it is not realistic to require good

[1] Note that computationally, this filtering (i.e. focusing) does not necessarily need to be implemented as a second step.

performance for both small and large faults. For example to require that the diagnosis system detects and uniquely isolates a very small bias is not realistic, but it may be realistic to require both good detection and isolation for large biases. Thus, the required performance of a diagnosis system need to be formulated differently for small and large faults respectively.

Formally, we start by partitioning the set Φ_b into two subsets Φ_b^{sig} and Φ_b^{insig}, representing *significant faults* and *insignificant faults* respectively. We will below use different performance requirements for these two sets. The idea of this partitioning is that Φ_b^{sig} contains the pdf's of those faults that are critical to detect and isolate. The set Φ_b^{insig} is then the pdf's of the faults that neither need to be detected or isolated, typically the smallest faults. The only requirement on these faults is that they should not lead to problems like "erroneous isolation" which can for example be seen as the event that a fault is present in only one component c_1 and at the same time the candidate sets tell that some other components are faulty but not c_1.

6.1 Performance Measures for Significant Faults

For each pdf belonging to Φ_b^{sig}, we use the following measures corresponding to coverage and false coverage probability respectively:

$$P(b \in C_F \mid f_b(z_\mathcal{T})) \tag{15}$$

$$P(b' \in C_F \mid f_b(z_\mathcal{T})) \quad b' \neq b \tag{16}$$

Still, the number of performance measures will typically be infinite. A solution to handle this is given later, together with the application example, in Section 9.

6.2 Performance Measure for Non-significant Faults

For the distributions belonging to Φ_b^{insig}, we use another performance measure based on the requirement that faults associated with Φ_b^{insig} should not lead to problems like erroneous isolation. The basic idea is that for modes in Φ_b^{insig}, we do not care about false coverage at all and we do not aim at coverage. Instead we aim only for something that we will call *sub-coverage*. Note that to not care about false coverage means that if b is the present mode, it is acceptable to also have other modes b' included in C_F.

The idea of *sub-coverage* is that we consider it fully acceptable to say that a component is non-faulty even though it is faulty. For example, if $b = [NF, F, NF, F]$ is the present mode and $z_\mathcal{T}$ has a distribution belonging to Φ_b^{insig}, it is acceptable if $[NF, F, NF, F] \notin C_F$ as long as $[NF, NF, NF, F]$, $[NF, F, NF, NF]$, or $[NF, NF, NF, NF]$ belong to C_F.

To formalize this, use ψ_i to denote the behavioral mode of the i:th component which means that b can be written as $b = [\psi_1, \psi_2, \ldots, \psi_n]$. Then let $\leq_O$ be a relation [2], defined on the set of system behavioral modes, such that $b' \leq_O b$, where $b' = [\psi_1', \psi_2', \ldots, \psi_n']$, if and only if $\forall i \in \{1, 2, \ldots n\} : \psi_i' = NF \vee \psi_i' = \psi_i$. By using this relation we replace the performance measure of coverage probability (15) with a measure that we call *sub-coverage probability*:

$$P(\exists \bar{b} \in C_F : \bar{b} \leq_O b \mid f_b(z_\mathcal{T})) \tag{17}$$

Note that the aim to make sub-coverage probability large includes the aim to make probability of "erroneous isolation" low.

[2] If system behavioral modes are represented by their sets of faulty components the relation $\leq_O$ is equivalent to the subset relation.

6.3 Bounds for Sub-Coverage

The aim now is to derive a useful bound for the probability of sub-coverage. We do this for the special case when the preference relation $\geq_p$ is such that $b' \geq_p b$ implies $b' \leq_O b$.

Theorem 2. If the preference relation $\geq_p$ is such that $b' \geq_p b$ implies $b' \leq_O b$, then for any $f_b(z_\mathcal{T}) \in \Phi_b$, it holds that

$$P(\exists \bar{b} \in C_F : \bar{b} \leq_O b \mid f_b(z_\mathcal{T})) \geq P(b \in C \mid f_b(z_\mathcal{T})) \tag{18}$$

Proof. If $b \in C$ then there is a mode b', where $b' \geq_p b$, and $b' \in C_F$. Since it holds that $b' \geq_p b$ implies $b' \leq_O b$, it follows that

$$\exists \bar{b} \in C_F : \bar{b} \leq_O b \tag{19}$$

Thus, we have proven that $b \in C$ implies (19). This fact means that (18) holds trivially. □

As seen this theorem shows that if we aim for coverage in C we get also sub-coverage.

7. COMPARISON OF FOCUSING PRINCIPLES

In this section we will compare the diagnosis system performance when using minimal and minimal cardinality diagnosis as focusing strategies and also the case without focusing. We use the performance measures defined in the previous section, i.e. coverage, false coverage, and sub-coverage. For sake of simplicity, we assume that (11) holds.

7.1 No Focusing

First, consider the strategy to not use focusing, i.e. $C_F = C$. Since we assume that (11) holds, the bound (13) with $B = \emptyset$ gives directly that $P(b \in C_F | f_b(z_\mathcal{T})) = 1$. Since $C_F \subseteq C$ also $P(b \in C | f_b(z_\mathcal{T})) = 1$, which implies, according to Theorem 2, that also $P(\exists \bar{b} \in C_F : \bar{b} \leq_O b \mid f_b(z_\mathcal{T})) = 1$. Thus both coverage and sub-coverage are guaranteed.

In general, false coverage can not be avoided. A typical example is if $[F, NF, NF]$ is the present mode. Then, assuming we have coverage, it holds that $[F, NF, NF] \in C$ but also that $[F, F, NF] \in C$ since it is typically not possible to construct a diagnostic test which responds to the mode $[F, NF, NF]$ but not to $[F, F, NF]$. Such a response would require that the second fault always compensates for the first one, something that is a rare situation in most real systems. Therefore, if b is the present mode, and we have coverage, all modes $\bar{b} \geq_O b$ will in the generic case be part of C_F. Thus, we can not avoid false coverage.

7.2 Focusing

We saw in the previous section that no focusing gives perfect performance with respect to coverage and sub-coverage, but very bad false coverage performance. The bad false coverage performance is the reason why focusing is used and we will in this section quantify how focusing improves the false coverage performance but also how the coverage performance is reduced if no special care is taken. We will later, in Section 7.3 and 7.4, see also that the sub-coverage performance may be severely affected depending on the actual focusing strategy chosen.

First consider the coverage probability. If **NF** is more preferred than any other mode, which should hold in any sensible focusing strategy, coverage in the case the present mode is **NF** is guaranteed from the bound (13) since the set B will be empty and we assume that (11) holds. For other modes, we do not get coverage automatically. When mode b is present, we need

the tests to respond in a way such that all modes $\bar{b} >_p b$ are eliminated from C_F. A sufficient condition to achieve coverage with high probability is obtained from the bound (13). This relation says that for each $\bar{b} >_p b$ it is sufficient to have one test that responds to b but not to $\bar{b}$ with high probability. Then the sum will be close to $|B|$ which implies that the bound becomes close to 1. Thus, the selection and design of a set of tests with this property for all significant faults is critical to obtain high coverage probability.

As said above, the only reason to use focusing is to lower the probability of false coverage. Given a mode b, consider the modes $\bar{b}$ for which it holds that $\bar{b} <_p b$ or $\bar{b} >_p b$. For these modes it holds that $b \in C_F$ implies $\bar{b} \notin C_F$. Therefore we have $P(\bar{b} \notin C_F|f_b(z_\mathcal{T})) \geq P(b \in C_F|f_b(z_\mathcal{T}))$. Thus, if we aim for high probability of coverage of b, which is of primary importance, we get also low false coverage probability of the pair $(\bar{b}, b)$.

Next, if $\bar{b} \not<_p b$ and $\bar{b} \not>_p b$, low false coverage probability can be guaranteed via the upper bound in (9) or the simplified bound (10). If the simplified bound is used, it tells us that a sufficient condition to get low false coverage probability is to, for each mode $\bar{b}$ where $\bar{b} \not<_p b$ and $\bar{b} \not>_p b$, have one test with $\bar{b} \in H_0^k$ and which responds with high probability when b is present.

7.3 Minimal Diagnoses

Now consider the case of focusing by means of the principle of *minimal diagnoses* [7]. This principle says that $\geq_p = \leq_O$. That means for example that if $[F, NF, NF] \in C$ and $[F, F, NF] \in C$, the mode $[F, NF, NF]$ is preferred and thus, $[F, F, NF] \notin C_F$. The underlying idea of this focusing principle is that if a diagnosis system says that mode $[F, NF, NF]$ is consistent with observations, there is no reason to believe that the a-priori much less probable mode $[F, F, NF]$ is the present mode.

All discussions in Section 7.2 regarding coverage and false coverage performance are valid for the case minimal diagnosis focusing. In addition we can note that, as a direct consequence of Theorem 2, the probability of sub-coverage is always greater than the coverage probability when minimal diagnoses focusing is used.

7.4 Minimal Cardinality Diagnoses

Next consider the focusing strategy *minimal cardinality*. This principle says that $b_1 \geq_p b_2$ if the number of faulty components in b_1 is less or equal to the number of faulty components in b_2. For example, $[F, NF, NF] >_p [NF, F, F]$. As in the case of minimal diagnosis focusing, all discussions in Section 7.2 regarding coverage and false coverage performance are valid for the case of minimal cardinality diagnosis focusing. However, there is an important difference regarding sub-coverage, something that is revealed by the following example. Assume that we have a diagnosis system with the following decision structure and that each test δ_k is designed to respond to the mode of a column if the row contains an X in the column.

	F1	**NF**	**F2**	**F3**	**F12**	**F23**	**F13**
δ_1	0	X	X	0	X	X	X
δ_2	0	X	0	X	X	X	X
δ_3	0	0	X	X	X	X	X

(20)

Assume the mode $\mathbf{F23}$ is present with an insignificant fault and because the fault is small, only tests T_1 and T_2 respond. This implies that $C = \{\mathbf{F1}, \mathbf{F12}, \mathbf{F23}, \mathbf{F13}\}$. Minimal cardinality focusing gives $C_F = \{\mathbf{F1}\}$. It is obvious that sub-coverage is not obtained. Note that in the case of minimal diagnosis focusing, sub-coverage is obtained (even coverage) since $C_F = \{\mathbf{F1}, \mathbf{F23}\}$.

The important conclusion of this study is that if an insignificant fault is present, we have no control of whether tests respond or not, and thus we can not guarantee any level of sub-coverage probability when using minimal cardinality focusing.

8. GUIDELINES FOR DESIGN OF DIAGNOSIS SYSTEMS

In Section 6 we have presented three fault-isolation performance-measures: coverage probability, false coverage probability, and sub-coverage probability. In this section we aim at giving some general design guidelines such that desired performances with respect to these three measures are obtained or maximized. First however we give some general presumptions as a starting point.

In Section 3.1 it was argued that lack of coverage can not be accepted in industrial applications. Therefore, but also to make our analysis tractable, we decide to aim for coverage probability one, i.e. $P(b \in C_F|f_b(z_\mathcal{T})) = 1$ for significant faults.

In Section 3.2 it was argued that false coverage is not as serious as lack of coverage. Therefore, and because we would often get an unsolvable problem if we would require false coverage with probability zero, we will not aim at $P(\bar{b} \in C_F|f_b(z_\mathcal{T})) = 0$ when $\bar{b} \neq b$ and the fault is significant. Instead we aim at $P(\bar{b} \in C_F|f_b(z_\mathcal{T})) \leq \epsilon$ where ϵ may be fixed or dependent on the pair $(\bar{b}, b)$.

We assume that the diagnosis system design starts with a default set of diagnostic tests where each test δ_k has a residual generator r_k and a set H_0^k. This situation is common for example if the diagnosis system design starts with a search for residual generators via structural analysis [8].

The design freedom then consists of: (i) selecting the rejection region, i.e. the threshold and possibly some residual filtering, of each test δ_k, (ii) from the default set select tests δ_k to be included in the diagnosis system, and (iii) to select the focusing strategy.

8.1 Selection of Rejection Region

A necessary requirement for coverage is that $P(b \in C|f_b(z_\mathcal{T})) = 1$ and from Theorem 1, it can be shown that a necessary and sufficient condition to achieve this is that the rejection region, for each diagnostic test δ_k, fulfills

$$P(\text{rej}_k|f_b(z_\mathcal{T})) = 0 \qquad \text{for all} \qquad b \in H_0^k \tag{21}$$

This rule is, as seen in Section 7.2, however not sufficient to obtain coverage in the case when focusing is used. When $\bar{b} \geq_p b$, coverage can only be guaranteed if we also have at least one test that responds to b but not to $\bar{b}$. Further, from (10) it is clear that also to obtain low false coverage probability, it is important to have tests that responds as much as possible to modes $b \notin H_0^k$. These facts means that we must follow the constraint (21) but in addition, it is in general advantageous to maximize the probability $P(\text{rej}_k|f_b(z_\mathcal{T}))$. This leads us to our first design guideline:

G1. For each diagnostic test δ_k, select the maximal rejection region such that $P(\text{rej}_k|f_b(z_\mathcal{T})) = 0$ for all modes $b \in H_0^k$ and all distributions $f_b(z_\mathcal{T}) \in \Phi_{\bar{b}}^{insig} \cup \Phi_{\bar{b}}^{sig}$.

8.2 Selection of Diagnostic Tests to Include

Following design guideline G1 is necessary to obtain coverage but as seen in Section 7.2 not sufficient if focusing is used. As was stated above, a sufficient condition is to, for each pair of

modes such that $\bar{b} >_p b$, have at least one test that responds to b with probability one but not to $\bar{b}$. From Section 7.2 it has already been concluded that if coverage of a b is secured, we only have to consider false coverage of modes $\bar{b}$ where $\bar{b} \not<_p b$ and $\bar{b} \not>_p b$. This leads us to our next design guideline:

G2. For each pair of modes $(\bar{b}, b)$, make sure that for all distributions $f_{\bar{b}} \in \Phi_{\bar{b}}^{insig} \cup \Phi_{\bar{b}}^{sig}$ and $f_b \in \Phi_b^{sig}$ there is, included in the diagnosis system, at least one test δ_k such that $\bar{b} \in H_0^k$, $P(\text{rej}_k | f_{\bar{b}}(z_\mathcal{T})) = 0$, and
   a) $P(\text{rej}_k | f_b(z_\mathcal{T})) = 1$ if $\bar{b} >_p b$
   b) $P(\text{rej}_k | f_b(z_\mathcal{T})) \geq 1 - \epsilon$ if $\bar{b} \not<_p b$ and $\bar{b} \not>_p b$

8.3 Selection of Focusing Strategy

Note that a consequence of the discussion in Section 7.1 is that fulfillment of guideline G2 is in general not possible if we don't use a focusing strategy. This implies that, of the three choices of no focusing, minimal diagnoses, and minimal cardinality diagnoses, we have to use minimal diagnoses or minimal cardinality diagnoses.

We have seen in Section 7 that the choice of focusing method affects the ability to obtain sub-coverage. Of the two choices left, i.e. minimal diagnoses and minimal cardinality diagnoses, minimal diagnosis is the best choice since it guarantees high sub-coverage probability when we have high coverage probability. This is our final design guideline:

G3. Use the focusing strategy *minimal diagnoses*.

8.4 Summarizing Theorem

We end this section by summarizing the discussion in a theorem.

Theorem 3. If guidelines G1, G2, and G3 are followed, we obtain a diagnosis system where:

a) $P(b \in C_F | f_b(z_\mathcal{T})) = 1$ for all $f_b(z_\mathcal{T}) \in \Phi_b^{sig}$ and for all b, i.e. coverage is guaranteed for all significant faults,
b) $P(\exists \bar{b} \leq_O b : \bar{b} \in C_F \,|\, f_b(z_\mathcal{T})) = 1$ for all $f_b(z_\mathcal{T}) \in \Phi_b^{insig}$ and for all b, i.e. sub-coverage is guaranteed for all insignificant faults,
c) $P(b' \in C_F \,|\, f_b(z_\mathcal{T})) \leq \epsilon$ for all $f_{\bar{b}} \in \Phi_{\bar{b}}^{insig} \cup \Phi_{\bar{b}}^{sig}$ and $f_b \in \Phi_b^{sig}$ and for all pairs $(\bar{b}, b)$, i.e. false coverage probability less than ϵ is guaranteed.

Further, no other choice of rejection region for each test gives strictly better performance in all measures of coverage, sub-coverage, or false coverage probability.

9. EXAMPLE

Consider a system with a pump P and two sensors S_1 and S_2. The angular velocity x of the pump is measured by sensor S_1. The angular velocity determines the output pressure which is measured by sensor S_2. The measurement signals are denoted y_1 and y_2 respectively. All three components are assumed to be either in a non-faulty NF or faulty mode F. The system behavioral modes are denoted by their faulty components, e.g. $\mathbf{S_1}$ means the mode where only the sensor S_1 is faulty.

Next, we assume that the following model is available:

$$P = NF \rightarrow u_a = u \tag{22a}$$
$$\dot{x} = f(x) + u_a \tag{22b}$$
$$S_1 = NF \rightarrow y_1 = x \tag{22c}$$
$$S_2 = NF \rightarrow y_2 = g(x) \tag{22d}$$
$$S_2 = F \rightarrow y_2 = g(x) + c \tag{22e}$$

where c is an unknown constant. Even though not written out explicitly we assume that all equations also are affected by noise terms with unspecified pdf's. Note that, and as will be shown below, it is for our purpose not important to know these unspecified pdf's explicitly.

According to our framework, the set of pdf's Φ_b, for each mode b, is assumed to be partitioned into two sets Φ_b^{sig} and Φ_b^{insig}. However, in this example, these sets are not specified explicitly. Instead we pick out, from each set Φ_b^{sig}, a pdf $f_b^*(z_\mathcal{T})$ that represents a *benchmark fault*. Then the benchmark fault is defined explicitly and we assume that the pdf $f_b^*(z_\mathcal{T})$ is representative for the whole set Φ_b^{sig} in the sense that for each $f_b(z_\mathcal{T}) \in \Phi_b^{sig}$ it holds that $P(\text{rej}_k | f_b(z_\mathcal{T})) \geq P(\text{rej}_k | f_b^*(z_\mathcal{T}))$ for all k.

It is assumed that only modes $\mathbf{P}$, $\mathbf{S_1}$, $\mathbf{S_2}$, and $\mathbf{S_1S_2}$ are important to detect and isolate and thus, only these are considered to have significant faults and consequently also benchmark faults. The benchmark fault for mode $\mathbf{P}$ is defined by replacing equation (22a) by $u_a = u + \Delta u_{\min}$, and the benchmark fault for mode $\mathbf{S_1}$ is defined by replacing equation (22c) by $y_1 = x + a_{\min}$. Further the benchmark fault for mode $\mathbf{S_2}$ is defined by $c = c_{\min}$. Finally, the benchmark fault for mode $\mathbf{S_1S_2}$ is the combination of the benchmark faults for $\mathbf{S_1}$ and $\mathbf{S_2}$.

Next, structural analysis, see [8], is used to find the equation sets that can be used to derive residual generators and their corresponding null hypotheses. The result is that 7 sets are found and the decision structure for potential tests δ_k, to be constructed from these equation sets found, is the following.

	equation set	**NF**	**P**	**S$_1$**	**S$_2$**	**S$_1$S$_2$**
δ_1	(22a), (22b), (22c)	0	X	X	0	X
δ_2	(22a), (22b), (22d)	0	X	0	X	X
δ_3	(22c), (22d)	0	0	X	X	X
δ_4	(22a), (22b), (22c), (22d)	0	X	X	X	X
δ_5	(22a), (22b), (22e)	0	X	0	0	0
δ_6	(22c), (22e)	0	0	X	0	X
δ_7	(22a), (22b), (22c), (22e)	0	X	X	0	X

(23)

In the decision structure above only modes which have significant faults are shown. All other multiple-fault modes have X:s only in their columns.

9.1 Diagnosis System Design

Now we have all the elements needed to start the design of the diagnosis system. By following guideline G3 we will use the focusing strategy minimal diagnoses. By using guideline G2 we will now describe how to, from the list of potential tests (23), select a subset of tests Δ to be included in the diagnosis system.

Given the focusing strategy and the significant faults considered, it follows that there is one requirement in guideline G2a for each pair in $R_a = \{(\mathbf{NF}, \mathbf{P}), (\mathbf{NF}, \mathbf{S_1}), (\mathbf{NF}, \mathbf{S_2}), (\mathbf{NF}, \mathbf{S_1S_2}), (\mathbf{S_1}, \mathbf{S_1S_2}), (\mathbf{S_2}, \mathbf{S_1S_2})\}$ and in guideline G2b, one for each pair in $R_b = \{(\mathbf{S_1}, \mathbf{S_2}), (\mathbf{S_1}, \mathbf{P}), (\mathbf{S_2}, \mathbf{P}), (\mathbf{S_2}, \mathbf{S_1}), (\mathbf{P}, \mathbf{S_1}), (\mathbf{P}, \mathbf{S_2}), (\mathbf{P}, \mathbf{S_1S_2}), (\mathbf{S_1S_2}, \mathbf{P})\}$.

To illustrate how to fulfill these requirements, consider the pair $(\mathbf{S_2}, \mathbf{S_1S_2}) \in R_a$. To fulfill guideline G2 for $(\mathbf{S_2}, \mathbf{S_1S_2})$ we need a test where $\mathbf{S_2} \in H_k^0$. Potential tests fulfilling this are tests δ_k indexed $\{1, 5, 6, 7\}$. Note that, since we intend to follow guideline G1, it will hold that $P(\text{rej}_k | f_{\mathbf{S_2}}(z_\mathcal{T})) = 0$ for any test δ_k, $k \in \{1, 5, 6, 7\}$, if included in the diagnosis system. If we choose to include δ_5, a consequence of fulfilling G1 is also that $P(\text{rej}_5 | f_{\mathbf{S_1S_2}}(z_\mathcal{T})) = 0$. This implies that, since we are looking for tests that fulfill G2a for the pair $(\mathbf{S_2}, \mathbf{S_1S_2})$, there are only the potential tests $\{1, 6, 7\}$ left. Thus to fulfill guideline G2a for

$(\mathbf{S_2}, \mathbf{S_1S_2})$ we would need at least one of the potential tests in $\pi_1 = \{1, 6, 7\}$ to be included in the diagnosis system.

For all other pairs in $R_a \cup R_b$, sets π_i of potential tests are obtained in the same way. A necessary requirement for a diagnosis system with tests Δ to fulfill G2, is that the set Δ has a non-empty intersection with all sets π_i.

By applying a minimal hitting set algorithm [7], we get that the minimal test sets are $\{1, 2, 3, 5\}$, $\{2, 3, 5, 6\}$, and $\{2, 3, 5, 7\}$. Hence a set of tests Δ included in a diagnosis system fulfilling G2 must necessarily be a superset of some of these minimal test sets. This is however not sufficient since both G2a and G2b specify requirements on $P(\text{rej}_k|f_b(z_\mathcal{T}))$ for all $f_b \in \Phi_b^{sig}$.

Assume that we decide to investigate if the minimal test set $\{1, 2, 3, 5\}$ fulfills the requirement on $P(\text{rej}_k|f_b^*(z_\mathcal{T}))$ for all pairs in $R_a \cup R_b$. For this set, all requirements on $P(\text{rej}_k|f_b^*(z_\mathcal{T}))$ specified by G2a and G2b correspond to non-zero entries in the following table.

	NF	**P**	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_1S_2}$
δ_1	0	p_1	p_4	0	1
δ_2	0	p_2	0	p_6	1
δ_3	0	0	p_5	p_7	p_8
δ_5	0	p_3	0	0	0

(24)

Then from guidelines G2a and G2b we can derive the requirements that $\max(p_1, p_2, p_3) = 1$, $\max(p_4, p_5) = 1$, $\max(p_6, p_7) = 1$, and $p_i > 1 - \epsilon$ for all $i = 3, \ldots 8$. The constant ϵ is the guaranteed false coverage probability that in this example is chosen as $\epsilon = 0.1$.

The next step is to construct residual generators for the selected equation sets and investigate if the requirements in (24) are achievable by filtering and thresholding of these residuals. Observer based residual generators are derived for $k = \{1, 2, 5\}$ and a static residual generator is derived using equation set 3 in (23). Then the pdf's $f_b^*(z_\mathcal{T})$ corresponding to the benchmark faults are estimated using data from the real process. These estimated pdf's are then used for selecting, by means of thresholding and filtering, the rejection region in accordance with G1.

Assume that there are thresholds for the residuals such that the following performance $P(\text{rej}_k|f_b^*(z_\mathcal{T}))$ for the benchmark faults has been confirmed:

	NF	**P**	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_1S_2}$
1	0	1	1	0	1
2	0	0.8	0	1	1
3	0	0	0.95	0.97	0.98
5	0	0.9	0	0	0

(25)

By using this matrix, the bounds for $P(b \in C|b')$ in Theorem 1, where b' corresponds to the rows and b to the columns, are:

	NF	**P**	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_1S_2}$
NF	1	0	0	0	0
P	0	1	[0 0.1]	0	[0 0.1]
S_1	0	[0 0.05]	1	0	0
S_2	0	[0 0.03]	0	1	0
S_1S_2	0	[0 0.02]	0	0	1

The interpretation of the first row is that, when the present mode is NF then $C_F = \{\mathbf{NF}\}$ with probability 1. In row 3, we can see that when $\mathbf{S_1}$ is the present mode then $\mathbf{S_1} \in C_F$ but $\mathbf{P}$ will also be included in C_F with a probability less than 0.05. No other modes will be included in C_F. All diagonal elements are 1, i.e. complete coverage of all significant faults have been obtained. All non-diagonal elements are less or equal to 0.1 and this means that the false coverage probability is less than 10%. In fact, the false coverage probability is better than the guaranteed 10% for all modes except for P.

10. CONCLUSIONS

The first contribution of the paper is the formalization of "fault isolation performance" in noisy and uncertain systems. For this we have used the established notion of *coverage* and *false coverage* from the field of statistics. Further it has been noted that a different performance criteria is needed for small faults, and we have therefore introduced the third performance measure *sub-coverage*. We have also derived formal relations describing the relationship between fault isolation performance and the null-hypotheses and rejection regions of the tests. Further, the intrinsic fault isolation performance of different AI-based fault isolation schemes has been evaluated and it has been concluded that the well known principle of *minimal cardinality diagnosis* gives a very bad performance for the case of small faults. Finally, based on the performance measure and investigations, we have developed some general design guidelines that, if followed, guarantee and maximize the fault isolation performance.

REFERENCES

[1] M. Blanke, M. Kinneart, J. Lunze, and M. Staroswiecki. *Diagnosis and Fault-Tolerant Control*. Springer-Verlag, 2003.

[2] G. Casella and R. L.Berger. *Statistical Inference*. Duxbury Press, 1990.

[3] M.-O. Cordier, P. Dague, F. Levy J. Montmain, M. Staroswiecki, and L. Trave-Massuyes. Possible conflicts: a compilation technique for consistency-based diagnosis approach from the artificial intelligence and automatic control perspectives. *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, 34(5):2163– 2177, 2004.

[4] J. J. Gertler. *Fault Detection and Diagnosis in Enginerering Systems*. Marcel Dekker, Inc., 1998.

[5] W. Hamscher, L. Console, and J. de Kleer, editors. *Readings in Model-Based Diagnosis*. Morgan Kaufmann Publishers, 1992.

[6] J. De Kleer, A. K. Mackworth, and R. Reiter. Characterizing diagnoses and systems. *Artificial Intelligence*, 56, 1992.

[7] J. De Kleer and B.C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[8] Mattias Krysander. *Design and Analysis of Diagnosis Systems Using Structural Methods*. PhD thesis, Linköpings universitet, June 2006.

[9] M. Nyberg. Automatic design of diagnosis systems with application to an automotive engine. *Control Engineering Practice*, 7(8):993–1005, 1999.

[10] M. Nyberg and M. Krysander. Combining AI, FDI, and statistical hypothesis-testing in a framework for diagnosis. In *Proceedings of IFAC Safeprocess'03*, Washington, USA, 2003.

[11] M. Nyberg and M. Krysander. Statistical properties and design criterions for AI-based fault isolation. Technical Report LiTH-ISY-R-2843, Dept. of Electrical Engineering Linköpings universitet, 2007. URL:http://www.vehicular.isy.liu.se/Publications/.

[12] R. J. Patton, P. M. Frank, and R.N. Clark. *Issues of Fault Diagnosis for Dynamic Systems*. Springer, 2000.

[13] S. Ploix, S. Touaf, and J. M. Flaus. A logical framework for isolation in fault diagnosis. In *Proceedings of IFAC Safeprocess'03*, Washington, USA, 2003.

[14] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[15] S. Tuhrim, J. Reggia, and S. Goodall. An experimental study of criteria for hypothesis plausibility. *Journal of Experimental & Theoretical Artificial Intelligence*, 3(2):129–144, 1991.