A comparative analysis of AI and control theory approaches to model-based diagnosis

M-O. Cordier INRA/IRISA, Rennes
P. Dague LIPN, Paris 13 University
M. Dumas CEA, Saclay
F. Lévy LIPN, Paris 13 University
J. Montmain EERIE, Nîmes
M. Staroswiecki LAIL, Lille University
L. Travé-Massuyès LAAS, Toulouse

(as part of the French IMALAIA group – imalaia@laas.fr.)

Abstract

Two distinct and parallel research communities have been working along the lines of the Model-Based Diagnosis approach: the FDI community and the DX community that have evolved in the fields of Automatic Control and Artificial Intelligence, respectively. This paper clarifies and links the concepts that underlie the FDI analytical redundancy approach and the DX logical approach. The formal match of the two approaches is demonstrated and the theoretical proof of their equivalence is provided under various assumptions.

1. Introduction

Diagnosis is an active research topic which can be approached from different perspectives according to the type of knowledge available. The so-called Model-Based Diagnosis (MBD) approach rests on the use of an explicit model of the system to be diagnosed. Two distinct and parallel research communities have been using the MBD approach. The Fault Detection and Isolation (FDI) community uses techniques from control theory and statistical analysis. It has now reached a mature state and a number of very good surveys exist in this field (Patton & Chen 1991; Frank 1996; Iserman 1997). The DX community emerged more recently, with foundations in the fields of Computer Science and Artificial Intelligence (Reiter 1987; de Kleer & Williams 1987; Hamscher, Console, & de Kleer 1992).

The goals of the IMALAIA group are to agree upon a common FDI/DX terminology, to identify similarities and complementarities in the FDI and DX methods, and to contribute towards a unifying framework, thus taking advantage of the synergy of techniques from the two communities.

This paper clarifies the link between *parity equations or analytical redundancy relations* (ARR for short) and *conflicts* by introducing the notion of *potential conflicts* or *ARR supports.* The formal match of the two approaches is thus shown. The FDI and DX approaches used for fault localization are then analyzed from the two perspectives. The *exoneration* and the *no-compensation* assumptions which are implicit in FDI are made clear, and the theoretical proof of equivalence of the two approaches is included, according to adopted assumptions. For the sake of clarity, the study is carried out in a pure consistency-based framework, i.e. without fault models.

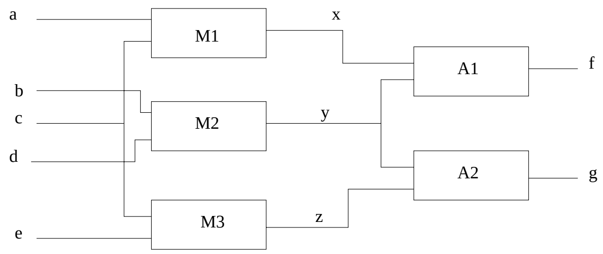

Figure 1: The system

The example that has been chosen to support the comparative analysis throughout the paper is the well-known system from (Davis 1984) composed of three multipliers M1, M2, M3 and two adders A1, A2 (see Figure 1). This choice and the fact that the system is assumed to operate in an ideal non-noisy and non-disturbed environment has been made on purpose to focus on the main features of each approach, without being overburdened neither with modeling details, nor with detection criteria. Let us emphasize that this discrete static example has been chosen for sake of clarity, but that the conclusions stemming from the comparison are quite general. In particular, both approaches can deal with continuous dynamic systems by basing the methods on differential or recurrent models. On the other side, the problems related to temporal diagnosis (Brusoni *et al.* 1998) involve many open issues in both approaches and are only evoked in the final discussion.

The paper is organized as follows. Sections 2 and 3 present the FDI analytical redundancy approach and the DX logical approach, respectively. Section 4 proposes a unified representation and proves the equivalence of the two approaches. This proof is given under specific assumptions corresponding to two classical cases which are the cases by default assumed in FDI and DX respectively. The general case and a more thorough analysis can be found in the long paper (Cordier *et al.* 2000). Finally, Section 5 discusses the results and outlines several interesting directions for future investigation.

2. Analytical redundancy-based diagnosis: the FDI approach

The FDI approach considers a dynamic model where the time evolution of the *measured* variables is a function of *input* variables, a (generally numeric) *fault vector* representing a deviation of parameters or inputs, and *noise* modeling uncertainty of data.

The *detection* problem is to decide whether the system is faulty or not. It rests on a partitioning of the fault vector space into two regions, the faulty and the correct ones. The decision can then be based on conditional probabilities of the pair (input,observation), under the two hypothesis – correct behavior, faulty behavior (Basseville & Nikiforov 1993). It can also be made by first estimating the fault vector before deciding what region it most likely belongs to (Iserman 1984).

Unknown initial state or inputs have often to be considered. They can be either estimated or eliminated. Algebraic or geometric elimination techniques are the essence of the parity space approach which produces Analytic Redundancy Relations (ARR) to get rid of the unknown initial state (Gertler 1991) or of unknown inputs (Staroswiecki & Comtet-Varga 2000).

The *isolation* task, which is in the following compared to the DX approach, is to decide which one among several fault hypotheses is most likely to be true. The FDI community most often uses geometric approaches, namely the so-called directional and structured residuals (Gertler & Singer 1990). Roughly, both techniques rest on a mapping of the triple (input,fault,noise) to a vector space of so called *evaluation forms* obtained through the elimination techniques previously considered. Structured residuals are such that each component of the residual vector depends only on a subset of the possible faults (this subset is called its structure). Directional residuals are structured residuals such that in the presence of each given fault, the residual vector gets colinear to a particular direction, which thus is the fault signature.

Let us now detail and exemplify some more how residuals are obtained and used. The behavioral model BM of a system is derived from its structure, which shows the links between its components *(structural model*), and the behavior model of each component.

Definition 2.1 The *system model* SM is defined as the *behavioral model* BM, i.e. the set of relations defining the system behavior, together with the *observation model* OM, i.e. the set of relations between the variables X of the system and the observed variables O acquired by the sensors.

Example Elementary components are the adders A1, A2, the multipliers M1, M2, M3 together with the set of sensors. The system model SM is hence given by the following:

BM:
RM1: $x = a \times c$ **RM2**: $y = b \times d$
RM3: $z = c \times e$ **RA1**: $f = x + y$
RA2: $g = y + z$
OM:
RSa: $a = a_{\text{obs}}$ **RSb**: $b = b_{\text{obs}}$
RSc: $c = c_{\text{obs}}$ **RSd**: $d = d_{\text{obs}}$
RSe: $e = e_{\text{obs}}$ **RSf**: $f = f_{\text{obs}}$
RSg: $g = g_{\text{obs}}$

Definition 2.2 A *diagnosis problem* is defined by the system model SM, a set of observations OBS assigning values to observed variables, and a set of faults F [1].

Example OBS = $\{a_{\text{obs}} = 2,\ b_{\text{obs}} = 2,\ c_{\text{obs}} = 3,\ d_{\text{obs}} = 3,\ e_{\text{obs}} = 2,\ f_{\text{obs}} = 10,\ g_{\text{obs}} = 12\}$.

The set of single faults is SF = $\{\text{F}_{\text{A1}}, \text{F}_{\text{A2}}, \text{F}_{\text{M1}}, \text{F}_{\text{M2}}, \text{F}_{\text{M3}}\}$, and the set of faults is F = 2^{SF}.

Definition 2.3 The *system structure* is defined through a binary application s: $SM \times V \rightarrow \{0,1\}$, where $V = X \bigcup O$ is the set of variables and *s(rel,v) = 1* if and only if v appears in relation *rel.*

Definition 2.4 An *analytical redundancy relation* (ARR) is a relation entailed by SM (and the components whose behavior model is used by this entailment are said to be *involved* in the ARR) which contains only observed variables, and which can therefore be evaluated from OBS. It is noted $r = 0$, where r is called the *residual* of the ARR. For a given OBS, the instantiation of the residual is noted *val(r,OBS)*, abbreviated as *val(r)* when not ambiguous. Thus, *val(r,OBS)=0* if the observations satisfy the ARR.

ARRs can be obtained from the system model by eliminating the unknown variables. This problem can be formalized in a graph theoretical framework, which comes down to the well-known problem of finding a complete matching w.r.t. the unknown variables X in the bipartite graph whose incidence matrix is the matrix associated to the application s. In this system structure matrix representation, a complete matching appears as a selection of one and only one entry per column, corresponding to an unknown variable, and per row, corresponding to a SM relation.

Example A complete matching leads to the following ARRs:

ARR1: $r_1 = 0$ where $r_1 \equiv f_{\text{obs}} - a_{\text{obs}} \times c_{\text{obs}} - b_{\text{obs}} \times d_{\text{obs}}$

ARR2: $r_2 = 0$ where $r_2 \equiv g_{\text{obs}} - b_{\text{obs}} \times d_{\text{obs}} - c_{\text{obs}} \times e_{\text{obs}}$

[1] In order to facilitate the comparison with DX, and without loss of generality, a fault can be seen as a set of faulty components.

The components involved in ARR1 (resp. ARR2) are A1, M1, M2 (resp. A2, M2, M3).

If we assume that the sensors are not faulty, then the ARRs can be rewritten as:
ARR1: $f - (a \times c + b \times d) = 0$
ARR2: $g - (b \times d + c \times e) = 0$

Let us call the ARRs that are obtained from a given complete matching *elementary ARRs.* Given a set of elementary ARRs, additional redundancy relations can be obtained by combining the elementary ones.

Example A third redundancy relation ARR3 can be obtained: ARR3: $f - g - a \times c + c \times e = 0$

The components involved in ARR3 are A1, A2, M1, M3 (notice that it is not the union of the components involved in ARR1 and ARR2).

Besides analytical redundancy relations, a fundamental concept in the FDI approach is that of *fault signature.*

Definition 2.5 Given a set R of n ARRs and a set F = $\{F_1, F_2, \ldots, F_m\}$ of m faults, the signature of a fault F_j is given by the binary vector $FS_j = [s_{1j}\ s_{2j} \ldots s_{nj}]^T$ in which s_{ij} is given by:
$R \times F \rightarrow \{0,1\}$
$(ARR_i, F_j) \mapsto s_{ij} = 1$ if some components involved in F_j are involved in ARR_i
$\mapsto s_{ij} = 0$ otherwise.

The interpretation of some s_{ij} being 0 is that the occurrence of the fault F_j does not affect ARR_i, meaning that *val*$(r_i) = 0$. The interpretation of some s_{ij} being equal to 1 is that the occurrence of the fault F_j is expected to affect ARR_i, meaning that *val*(r_i) is expected to be different from 0.

Definition 2.6 Given a set R of n ARRs, the signatures of a set F of m faults all put together constitute the so-called *signature matrix.*

Example The signature matrix for the set of single faults corresponding to components A1, A2, M1, M2 and M3, respectively, is given by:

	F_{A1}	F_{A2}	F_{M1}	F_{M2}	F_{M3}
ARR1	1	0	1	1	0
ARR2	0	1	0	1	1
ARR3	1	1	1	0	1

The case of multiple faults can be dealt with by expanding the number of columns of the signature matrix, leading to a total number of 2^m-1 columns with m the number of single faults, if all the possible multiple faults are considered. Let F_J be a multiple fault corresponding to the occurrence of k single faults $F_{j1}, \ldots, F_{jk}$, then the entries of the signature vector of F_J are given by:
$s_{ij} = 0$ if $s_{i\ j1} = \ldots = s_{i\ jk} = 0$
$s_{ij} = 1$ if $\exists l\ 1 \leq l \leq k$ such that $s_{i\ jl} = 1$

Example Extending the matrix above, the 26 additional columns have a $[1, 1, 1]^T$ signature, except for $F_{\{A1, M1\}}$ which has a $[1, 0, 1]^T$ signature, and for $F_{\{A2, M3\}}$ which has a $[0, 1, 1]^T$ signature.

The diagnostic sets in the FDI approach are given in terms of the faults accounted for in the signature matrix. The generation of the diagnostic sets is based on a column interpretation of the signature matrix and consists in comparing the *observation signature* with the fault signatures. This comparison is stated as a decision-making problem.

Definition 2.7 The signature of a given observation OBS is a binary vector $OS = [OS_1, \ldots, OS_n]^T$ where $OS_i = 0$ if and only if *val*$(r_i, OBS) = 0$.

The first step (the *detection* task) is to build the observation signature, i.e. to decide whether a residual value is zero or not, in the presence of noises and disturbances. This problem has been thoroughly investigated within the FDI community, much less within the DX community. It is generally stated as a statistical decision-making problem, making use of the available noise and disturbance models.

Example With OBS as above, $OS = [1, 0, 1]^T$. In the case $f = 10$ and $g = 10$, $OS = [1, 1, 0]^T$ and in the case $f = 10$ and $g = 14$, $OS = [1, 1, 1]^T$.

The second step (the *isolation* task) is to actually compare the observation signature with the fault signatures. A solution to this decision-making problem is to define a *consistency criterion* as follows:

Definition 2.8 An observation signature $OS = [OS_1, \ldots, OS_n]^T$ is consistent with a fault signature $FS_j = [s_{1j}, \ldots, s_{nj}]^T$ if and only if $OS_i = s_{ij}$ for all i.

Definition 2.9 The *diagnostic sets* are given by the faults whose signatures are consistent with the observation signature.

Example The following results are obtained for different observation signatures:
$OS = [1, 0, 1]^T \Leftrightarrow F_{A1}$ or F_{M1} or $F_{\{A1, M1\}}$
$OS = [1, 1, 0]^T \Leftrightarrow F_{M2}$
$OS = [1, 1, 1]^T \Leftrightarrow$ any multiple fault except $F_{\{A1, M1\}}$ and $F_{\{A2, M3\}}$

Note that the FDI community generally uses a *similarity-based consistency criterion* arising from the definition of a distance rather than the equality-based criterion defined above.

3. Logical-based diagnosis: the DX approach

Reiter (Reiter 1987) proposed a logical theory of diagnosis. This approach, also referred to as consistency-based diagnosis, was later extended and formalized in (de Kleer, Mackworth, & Reiter 1992). In the following we refer to the basic definitions of (Reiter 1987) without considering posterior extensions and refinements. The description of the behavior of the system is component-oriented and rests on first-order logic.

Definition 3.1 A *system model* is a pair (SD, COMPS) where SD, the *system description,* is a set of first order logic formulas with equality and COMPS, the components of the system, is a finite set of constants. SD uses a distinguished predicate AB, interpreted to mean abnormal. $\neg$AB(c) with c belonging to COMPS hence describes the case where the component c is behaving correctly.

Example COMPS = {A1, A2, M1, M2, M3}
SD = { ADD(x) $\wedge \neg$AB(x) $\Rightarrow$ Output(x) = Input1(x) + Input2(x),
MULT(x) $\wedge \neg$AB(x) $\Rightarrow$ Output(x) = Input1(x) $\times$ Input2(x),
ADD(A1), ADD(A2), MULT(M1), MULT(M2), MULT(M3),
Output(M1) = Input1(A1), Output(M2) = Input2(A1), Output(M2) = Input1(A2), Output(M3) = Input2(A2), Input2(M1) = Input1(M3) }

Let us note one point which differs somewhat from the description of the system in the FDI approach: with the distinguished predicate AB it is possible to make explicit the fact that a formula in SD describes the normal behavior of a given component. The description can easily be extended to include faulty behaviors.

A diagnosis problem results from the discrepancy between the normal behavior of a system as described by the system model and a set of observations.

Definition 3.2 A set of observations OBS is a set of first-order formulas.

Example An example of observations for our system is OBS = {Input1(M1) = 2, Input2(M1) = 3, Input1(M2) = 2, Input2(M2) = 3, Input2(M3) = 2, Output(A1) = 10, Output(A2) = 12}.

Definition 3.3 A *diagnosis problem* is a triple (SD, COMPS, OBS) where (SD, COMPS) is a system model and OBS a set of observations.

A diagnosis is a conjecture that certain components of the system are behaving abnormally. This conjecture has to be consistent with what is known about the system and with the observations.

Definition 3.4 A *diagnosis* for (SD, COMPS, OBS) is a set of components $\Delta \subseteq$ COMPS such that SD $\bigcup$ OBS $\bigcup$ {AB(c) | c $\in \Delta$ } $\bigcup$ { $\neg$AB(c) | c $\in$ COMPS – Δ} is satisfiable. A *minimal diagnosis* is a diagnosis Δ such that $\forall \Delta' \subset \Delta$, Δ' is not a diagnosis.

Following the principle of parsimony, minimal diagnoses are often the preferred ones. For the sake of simplicity, we will limit ourselves to minimal diagnoses. A method based upon the concept of conflict set has been proposed in (Reiter 1987) to generate minimal diagnoses and is at the basis of most of implemented DX algorithms.

Definition 3.5 An *R-conflict* for (SD, COMPS, OBS) is a set of components C = {c1, . . . , ck} $\subseteq$ COMPS such that SD $\bigcup$ OBS $\bigcup$ {$\neg$AB(c) | c $\in$ C} is inconsistent. A *minimal R-conflict* is an R-conflict which does not include any R-conflict.

An R-conflict can be interpreted as follows: one at least of the components in the R-conflict is faulty in order to account for the observations.

Example The system with the observations as seen above has the following minimal R-conflicts: {A1, M1, M2} and {A1, A2, M1, M3} due to the abnormal value of 10 for f. In the case f = 10 and g = 10, the two minimal R-conflicts are: {A1, M1, M2} and {A2, M2, M3}. In the case f = 10 and g = 14, there are three minimal R-conflicts: {A1, M1, M2}, {A2, M2, M3} and {A1, A2, M1, M3}.

Using these minimal R-conflicts, it is possible to give a characterization of minimal diagnoses which provides a basis for computing them (Reiter 1987).

Proposition 3.1 Δ is a minimal diagnosis for (SD, COMPS, OBS) if and only if Δ is a minimal hitting set [2] for the collection of (minimal) R-conflicts for (SD, COMPS, OBS).

Example With f = 10 and g = 12, there are four minimal diagnoses given by the minimal hitting sets for {{A1, M1, M2}, {A1, A2, M1, M3}} which are: Δ1 = {A1}, Δ2 = {M1}, Δ3 = {A2, M2}, Δ4 = {M2, M3}.

With f = 10 and g = 10, there are five minimal diagnoses given by the minimal hitting sets for {{A1, M1, M2}, {A2, M2, M3}} which are: Δ1 = {M2}, Δ2 = {A1, A2}, Δ3 = {A1, M3}, Δ4 = {A2, M1}, Δ5 = {M1, M3}.

With f = 10 and g = 14, there are eight minimal diagnoses given by the minimal hitting sets for {{A1, M1, M2}, {A2, M2, M3}, {A1, A2, M1, M3}} which are: Δ1 = {A1, A2}, Δ2 = {A1, M2}, Δ3 = {A1, M3}, Δ4 = {A2, M1}, Δ5 = {A2, M2}, Δ6 = {M1, M2}, Δ7 = {M1, M3}, Δ8 = {M2, M3}.

4. Unified framework for the DX and FDI approaches

ARRs vs R-conflicts

In both approaches, diagnosis is triggered when discrepancies occur between the modeled (correct) behavior and the observations (OBS). In the ARR framework, discrepancies come from ARRs which are not satisfied by OBS. In DX, discrepancies allow the identification of R-conflicts, where an R-conflict is a set of components the correctness of which supports a discrepancy. An analogous concept can be defined in FDI.

Definition 4.1 The *support* of an ARR is the set of components involved in this ARR, i.e. columns of the signature matrix with a non zero element in the row

[2] A hitting set for a collection of sets is a set that intersects any set of the collection.

corresponding to this ARR. It is also called a *potential R-conflict*. This name is justified by the following result.

Proposition 4.1 Let OBS be a set of observations for a system modeled by SM (resp. SD). There is an identity between the set of minimal R-conflicts for OBS and the set of minimal potential R-conflicts associated to the ARRs which are not satisfied by OBS (see proof in (Cordier *et al.* 2000)).

Example The potential R-conflicts are: C1 = {A1, M1, M2} (support of ARR1), C2 = {A2, M2, M3} (support of ARR2) and C3 = {A1, A2, M1, M3} (support of ARR3).
With f = 10 and g = 12, ARR1 and ARR3 are not satisfied, which gives rise to the minimal R-conflicts C1 and C3.
With f = 10 and g = 10, ARR1 and ARR2 are not satisfied, which gives rise to the minimal R-conflicts C1 and C2.
With f = 10 and g = 14, ARR1, ARR2 and ARR3 are not satisfied, which gives rise to the minimal R-conflicts C1, C2 and C3.

Let us now analyze the relationship between potential R-conflicts and R-conflicts. From the computational point of view, the main difference between the FDI and DX approaches is that in FDI most of the work is done off-line. Using just the knowledge of observed variables, i.e. sensor locations, modeling knowledge is compiled: ARRs are obtained by combining model constraints and eliminating unobserved variables. The only thing that has to be done on-line, i.e. when a given OBS is acquired, is to compute the falsity value (w.r.t. OBS) of each ARR and to compare the observation signature obtained with the fault signatures. In terms of R-conflicts, this means that potential R-conflicts are compiled. This avoids any propagation: for any OBS, R-conflicts are exactly those potential R-conflicts which are supports of those ARRs which are not satisfied by OBS, so they are directly obtained from the detection task. Notice that such a compilation has already been used in the DX approach for a continuous dynamic system, the Monostable circuit (Loiez & Taillibert 1997; Loiez 1997).

The matrix framework

The FDI approach uses the signature matrix crossing ARRs in rows and sets of components in columns. It was shown in section 2 that, given an observation OBS, diagnosis is achieved by identifying those columns which are identical (or closest w.r.t. a distance function) to the observation signature column.

In the DX approach, it has been seen in section 3 that minimal diagnoses are obtained as minimal hitting sets of the collection of (OBS-) R-conflicts. From proposition 4.1 above, such R-conflicts can be viewed as the supports of those ARRs which are not satisfied by OBS, i.e. by looking at the corresponding set of rows I. A minimal hitting set of the collection of R-conflicts can thus be viewed as a minimal set J of singleton columns such that each of the rows of I intersects at least one column of J (i.e. has a non zero element in this column).

It is thus quite natural to adopt this matrix framework as a formal basis on which to compare the two approaches. The following notations are used:

- Let R = {$\mathrm{ARR_i}$ / i = 1 ...n} be the set of ARRs and COMPS = {$\mathrm{C_j}$ / j = 1 ...m} the set of components of the system. FS = $[\mathrm{s_{ij}}]_{\mathrm{i = 1...n,\ j = 1...m}}$ is the signature matrix. The $\mathrm{j^{th}}$ column of FS is the signature of a fault in $\mathrm{C_j}$ and is noted $\mathrm{FS_j}$. For J = {$\mathrm{j_1}$, ...,$\mathrm{j_k}$ } ⊆ {1,...,m}, let us note $\mathrm{C_J}$ the subset {$\mathrm{C_j}$ / j ∈ J}, and $\mathrm{s_{iJ}}$ the element of the extended matrix FS at line i and column J.
- Any observation OBS splits the set R into two subsets:
  - the subset $\mathrm{R_{false}}$ of ARRs it is inconsistent with, i.e. $\mathrm{R_{false}}$ = {$\mathrm{ARR_i} \equiv (r_i = 0)$ / $val(r_i, OBS) \neq 0$}.
  - the subset $\mathrm{R_{true}}$ = ARR – $\mathrm{R_{false}}$ of ARRs it is consistent with, i.e. $\mathrm{R_{true}}$ = {$\mathrm{ARR_i} \equiv (r_i = 0)$ / $val(ri, OBS) = 0$}.

  OBS is thus described through its signature OS, which is the binary column vector defined by: for all i = 1 ...n, $\mathrm{OS_i}$ = 1 if $\mathrm{ARR_i} \in \mathrm{R_{false}}$ and $\mathrm{OS_i}$ = 0 if $\mathrm{ARR_i} \in \mathrm{R_{true}}$.

The FDI theory compares the observation signature to the fault signatures whereas DX considers separately each line corresponding to an ARR in $\mathrm{R_{false}}$, isolating R-conflicts before searching for a common explanation. In the following, these approaches are called *column view* and *line view* respectively.

Exoneration and no-compensation assumptions

The originality and the power of both the FDI and DX approaches result from the fact that they are based only on the correct behavior of the components: no model of faulty behavior is needed. Nevertheless, different assumptions concerning the manifestations of the faults through observations are adopted by default by each approach, leading to different computations of the diagnoses, which explains the different results obtained on the example. These assumptions concern: 1) the manifestations of the faults through observations and 2) the case of simultaneous faults and of their interaction.

In addition to the obvious fact that a fault cannot affect an ARR in which it is not involved, which is the direct form of the reasoning used in DX, the idea used in FDI is that a fault necessarily manifests itself by affecting the ARRs in which it is involved, causing them not to be satisfied by any given OBS. Hence not only, as in DX, is any column involved in a not satisfied row a fault candidate, but also any column involved in a satisfied ARR is implicitly exonerated (satisfied rows are thus also used in the reasoning). In fact this result is not sound but rests on an exoneration assumption which is implicitly made in the FDI approach and has to be considered explicitly in order to compare the FDI approach with the DX approach.

Definition 4.2 (ARR-based exoneration assumption) A set of faulty components necessarily shows its faulty behavior, i.e. causes any ARR in which it is involved not to be satisfied by any given OBS. Or, equivalently, given OBS, any set of components involved in a satisfied ARR is exonerated, i.e. each component of its support is considered to be behaving correctly.

Note that this general exoneration assumption is made up of 1) a *single fault exoneration assumption* (each individual component shows its faulty behavior) and 2) a *no-compensation assumption* (the individual effects of faulty components never compensate each other).

From the matrix viewpoint, the fact that $\mathrm{ARR_i}$ exonerates $\mathrm{C_j}$ will appear as usual (cf. section 2) in FS as $\mathrm{s_{ij}} = 1$, whereas we have chosen to represent the fact that $\mathrm{C_j}$ is in the support of $\mathrm{ARR_i}$ but that the exoneration is not assumed by $\mathrm{s_{ij}} = \mathrm{X}$. The elements of FS can thus take their values in {0,1}, {0,X} or {0,X,1}. The semantics of $\mathrm{s_{ij}} = \mathrm{X}$ is: a fault in $\mathrm{C_j}$ can explain why $\mathrm{ARR_i}$ is not satisfied, but $\mathrm{ARR_i}$ may happen to be satisfied even when $\mathrm{C_j}$ is faulty. The semantics of $\mathrm{s_{ij}} = 1$ is: a fault in $\mathrm{C_j}$ forces $\mathrm{ARR_i}$ not to be satisfied (hence if $\mathrm{ARR_i}$ is satisfied then $\mathrm{C_j}$ is not faulty – which explains the term "exoneration"). The generalized use of an exoneration assumption for each component will be called the *exoneration and no-compensation case* and corresponds to the assumption by default in the FDI approach, while the total lack of exoneration will be called the *no-exoneration and compensation case* and corresponds to the assumption by default in the DX approach.

Equivalence in the exoneration and no-compensation case

In this case, fault signatures involve only 0 and 1.

As seen in section 2, the signature of the column $\mathrm{C_J}$ of the extended matrix is given by the following *fault interaction law* which expresses the no-compensation assumption:

$$\mathrm{s_{iJ}} = \sup\{\mathrm{s_{ij}} \;/\; \mathrm{j} \in \mathrm{J}\} \text{ for the order } 0 < 1 \qquad \text{(FIenc)}$$

We define $\mathrm{Support(ARR_i)} = \{\mathrm{C_J} \;/\; \mathrm{s_{iJ}} = 1\}$ and $\mathrm{Scope(C_J\,)} = \{\mathrm{ARR_i} \;/\; \mathrm{s_{iJ}} = 1\}$.

The column view searches for a perfect match of a fault signature with the observation signature. A set $\mathrm{C_J}$ is then a possible diagnosis if and only if:

$$\mathrm{R_{false}} = \mathrm{Scope(C_J)} \qquad \text{(CVenc)}$$

The line view is that possible diagnoses are subsets $\mathrm{C_J}$ of COMPS such that:

$$\begin{array}{l} \forall \mathrm{i}\ (\mathrm{ARR_i} \in \mathrm{R_{false}} \Rightarrow \\ \quad \exists \mathrm{j} \in \mathrm{J},\ \mathrm{C_j} \in \mathrm{Support(ARR_i)}) \\ \qquad\qquad \wedge \\ \forall \mathrm{i}\ (\mathrm{ARR_i} \in \mathrm{R_{true}} \Rightarrow \\ \quad \forall \mathrm{j} \in \mathrm{J},\ \mathrm{C_j} \in \mathrm{COMPS} - \mathrm{Support(ARR_i)}) \end{array} \qquad \text{(LVenc)}$$

Due to (FIenc) this is equivalent to:

$$\forall \mathrm{i}\ (\mathrm{ARR_i} \in \mathrm{R_{false}} \Leftrightarrow \mathrm{C_J} \in \mathrm{Support(ARR_i\,)})$$

which is itself equivalent to (CVenc), which proves the equivalence of the column and line views.

Example This equivalence is illustrated in the example.

With f = 10 and g = 12, i.e. observation signature (1,0,1), there are 2 minimal single fault diagnoses {A1} and {M1} and one superset diagnosis {A1, M1} (the components A2, M2 and M3 are exonerated as members of the support of the satisfied ARR2).

With f = 10 and g = 10, i.e. observation signature (1,1,0), the only diagnosis is {M2} (the components A1, A2, M1 and M3 are exonerated as members of the support of the satisfied ARR3).

With f = 10 and g = 14, i.e. observation signature (1,1,1), there are 8 minimal double fault diagnoses (those found in section 3) and 16 superset diagnoses (exoneration plays no role here).

Equivalence in the no-exoneration and compensation case

In this case, which is the common one in DX, fault signatures involve only 0 and X, and X matches both 0 and 1.

From the semantics of X seen in , it results that columns of the extended matrix are built according to the following rule: a multiple fault can explain that a given ARR is not satisfied if and only if at least one of its faults can explain it, i.e. several faults never produce more than the combination of their separate effects; on the other hand, it is admitted that a faulty component does not necessarily affect an ARR in which it is involved (single fault no-exoneration) and that several faults may always compensate each other (compensation), resulting in a satisfied ARR. The *fault interaction law* can thus be stated as:

$$\mathrm{s_{iJ}} = \sup\{\mathrm{s_{ij}} \mid \mathrm{j} \in \mathrm{J}\} \text{ for the order } 0 < \mathrm{X} \qquad \text{(FInec)}$$

We define $\mathrm{WeakSupport(ARR_i)} = \{\mathrm{C_J} \mid \mathrm{s_{iJ}} \neq 0\}$ and $\mathrm{WeakScope(C_J)} = \{\mathrm{ARR_i} \mid \mathrm{s_{iJ}} \neq 0\}$.

In the column view, $\mathrm{C_J}$ is a possible diagnosis if and only if:

$$\mathrm{R_{false}} \subseteq \mathrm{WeakScope(C_J\,)} \qquad \text{(CVnec)}$$

In the line view the diagnoses are the sets $\mathrm{C_J}$ such that:

$$\begin{array}{l} \forall \mathrm{i}\ (\mathrm{ARR_i} \in \mathrm{R_{false}} \Rightarrow \\ \qquad \exists \mathrm{j} \in \mathrm{J},\ \mathrm{C_j} \in \mathrm{WeakSupport(ARR_i)}) \end{array} \qquad \text{(LVnec)}$$

Due to (FInec), this translates to:

$$\forall \mathrm{i}\ (\mathrm{ARR_i} \in \mathrm{R_{false}} \Rightarrow \mathrm{C_J} \in \mathrm{WeakSupport(ARR_i\,)})$$

which in turn is the same as $\mathrm{R_{false}} \subseteq \mathrm{WeakScope(C_J)}$, i.e. (CVnec). This proves the equivalence of diagnoses.

Example The extended signature matrix is obtained from the usual one (see section 2) by replacing each 1 by X.

With f = 10 and g = 12, i.e. observation signature (1,0,1), there are 4 minimal diagnoses: the 2 single fault diagnoses {A1} and {M1} and the 2 double fault diagnoses {A2, M2} and {M2, M3}, and 22 superset diagnoses.

With f = 10 and g = 10, i.e. observation signature (1,1,0), there are 5 minimal diagnoses: the single fault diagnosis {M2} and the 4 double fault diagnoses {A1, A2}, {A1, M3}, {A2, M1} and {M1, M3}, and 20 superset diagnoses.

With f = 10 and g = 14, i.e. observation signature (1,1,1), the results are the same that in the exoneration case above.

Notice that, in the three cases of observation, single faults are identical to those obtained with exoneration assumption. This is because the single fault exoneration assumption is licit in the case of the example. The reason is that all component models are invertible, i.e. the value of each port is functional w.r.t. the values of the other ports. When this is not the case, it is easy to find examples where this assumption is not justified: it is sufficient to use and/or gates in place of adders/multipliers. For example, if the output (supposed not observable) of a component is connected to one input of an OR gate, whose second input is 1 and ouput is also 1, any ARR involving this component and the OR gate will be satisfied whatever the value of the output of the component is, i.e. whatever the component is faulty or not. This is why the single fault exoneration assumption adopted by default in FDI fails for particular devices with limited observability.

Concerning multiple faults, notice that in the two observation cases f = 10, g = 12 and f = 10, g = 10, all multiple faults (except {A1, M1} in the first case) discovered by the DX approach are wrongly ruled out by the FDI approach. This is because the no-compensation assumption adopted by default in FDI fails for particular cases. For example, with f = 10 and g = 12, the double fault {A2, M2} corresponds to the case where M2 behaves as $2 \times 3 = 4$ and A2 as $4 + 6 = 12$ and the double fault {M2, M3} to the case where M2 behaves as $2 \times 3 = 4$ and M3 as $3 \times 2 = 8$, introducing each time a compensation at the level of the output g, which is correct, resulting in ARR2 being satisfied and thus in the components A2, M2 and M3 of its support being wrongly exonerated by FDI. In the present example, such compensation cases appear to be exceptional due to the potentially infinite set of possible values for each variable, but this is not always the case. For example, in case of more discrete systems, i.e. with few possible values for each variable (such as boolean circuits), or in case of imprecise observation and/or rough (qualitative) models, it will happen frequently that multiple faults may compensate each other (e.g., in the sign algebra, a plus and a minus always compensate each other). It remains true that the more numeric and continuous is the model (which is the usual case in FDI) and the more precise is the observation, then the more licit is the no-compensation assumption: in this context, ARR-based no-compensation hypothesis is valid for almost every fault.

The general case

It is now simple to provide an extension of the framework which allows three-valued fault signatures, involving 0, X and 1. In this case, exoneration applies to some components w.r.t. some ARRs, but not to all. Equivalence can be proved in the same way as above (Cordier *et al.* 2000).

5. Conclusion and prospects

The first goal of FDI was fault detection and associated decision procedures. Its main interest was to offer sophisticated techniques so as to combine observations such as observers and filters. DX, on the other hand, aimed at localization by recognizing subsets of the system description that conflicted with the observation. Our study proves that a significant part of the two theories fits into a common framework which allows a precise comparison. When they adopt the same hypotheses with respect to how faults manifest themselves, FDI and DX views agree on diagnoses. This opens the possibility of a fruitful cooperation between these two diagnostic approaches, getting the best from each one: compiling modeling knowledge under ARRs form according to sensor locations before any observation has been made, which is the main advantage of the FDI approach ; and computing at the same time potential R-conflicts (supports of ARRs) to give rise, given an OBS, to R-conflicts on which the diagnoses generation is based, ensuring soundness of the diagnoses obtained w.r.t. to explicit assumptions about exoneration and compensation, which is the main advantage of the DX approach.

It is important to notice that the equivalence between the two approaches is obtained either by importing in DX the exoneration and no-compensation (enc) assumption implicitly used in FDI or by importing in FDI the no-exoneration (nec) assumption used by default in DX. As (nec) never eliminates wrongly diagnoses contrary to (enc), our equivalence results allow FDI approach to use (nec) in order to ensure soudness. Nevertheless, the concept of exoneration can be useful to rule out improbable diagnoses. This concept has been indeed introduced in DX, but expressed at the component model level instead of at the ARR level, which guarantees soundness. This is done by assuming that, if the correct behavior model of a component is satisfied by OBS, then this component behaves correctly in the context given by OBS, i.e. by modeling components behavior with bi-conditionals (Raiman 1992). In (Cordier *et al.* 2000) this model-based exoneration (mbe), which is proved to be weaker than (enc) in the single fault case, is thoroughly compared with (enc). An analog of the proposition 4.1, which relates minimal alibis, i.e. defined Horn AB-clauses entailed by SD $\bigcup$ OBS, with supports of ARRs satisfied by OBS, allows one to prove that any FDI diagnosis with (enc) is a DX diagnosis with (mbe) when SD $\bigcup$ OBS is Horn (but the converse is false). Then the comparison is made between (mbe) and what turns out to be the closest assumption in the FDI framework, i.e. fault exoneration and multiple fault compensation (ec): most of the time the diagnoses obtained are identical (this is the case for the example considered here) but this is not always true.

Some points need future investigation.

There is presently no equivalent in DX of the notion of noise and disturbance, because discrete state models that were originally studied by DX are robust by nature. Using in the DX framework the work accomplished by the FDI community in this field, in particular to perform the fault detection task, is thus needed for real applications.

Conversely, in the consistency-based extended framework, DX makes a systematic use of fault models, whose counterpart in FDI (real fault models are rarely used in FDI because they are difficult to obtain at a numeric level) can be found in assumptions about the additive or multiplicative deviations which model the faults. Fault models have been left out of the framework of the present paper. The comparison has thus to be extended to such fault models, looking in particular for an analog in FDI of DX conflicts which are not Horn AB-clauses.

The conclusions of this work remain valid in case of temporal sequence of observations when the faults are present from the beginning and do not evolve along time. Because, in this case, considering a sequence of observation does not modify the framework: more observation signatures from one hand and more conflicts from the other hand allow in a same way diagnoses to be refined, by reasoning on each snapshot of the system (state-based approach). Conversely, the incremental diagnosis problem (i.e. when faults occur and evolve along time, which is the case within the supervision task) is still open on each side: dealing with dynamic residuals and temporal signatures on one side and with simulation-based approach (vs. state-based approach (Struss 1997)) on the other side.

Further studies are needed to integrate these aspects, which would be beneficial to both communities.

References

Basseville, M., and Nikiforov, I. 1993. *Detection of abrupt changes, Theory and applications*. Information and System Sciences Series. Prentice Hall.

Brusoni, V.; Console, L.; Terenziani, P.; and Dupré, D. T. 1998. A spectrum of definitions for temporal model-based diagnosis. *Artificial Intelligence* 102(1):39–79.

Cordier, M.-O.; Dague, P.; Dumas, M.; Lévy, F.; Montmain, J.; Staroswiecki, M.; and Travé-Massuyès, L. 2000. Conflicts versus analytical redundancy relations. to be submitted.

Davis, R. 1984. Diagnostic reasoning based on structure and behavior. *Artificial Intelligence* 24:347–410.

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

de Kleer, J.; Mackworth, A.; and Reiter, R. 1992. Characterizing diagnoses and systems. *Artificial Intelligence* 56(2-3):197–222.

Frank, P. 1996. Analytical and qualitative model-based fault diagnosis – a survey and some new results. *European Journal of Control* 2:6–28.

Gertler, J. J., and Singer, D. 1990. A new structural framework for parity space equation based failure detection and isolation. *Automatica* 26:381–388.

Gertler, J. 1991. Analytical redundancy methods in fault detection and isolation, survey and synthesis. In *IFAC Safeprocess'91*, volume 1, 9–21.

Hamscher, W.; Console, L.; and de Kleer, J., eds. 1992. *Readings in Model-Based Diagnosis*. San Mateo, CA: Morgan Kaufmann.

Iserman, R. 1984. Process fault detection based on modeling and estimation methods - a survey. *Automatica* 20(4):387–404.

Iserman, R. 1997. Supervision, fault detection and fault-diagnosis methods – an introduction. *Control Engineering Practice* 5(5):639–652.

Loiez, E., and Taillibert, P. 1997. Polynomial temporal band sequences for analog diagnosis. In *15th International Joint Conference on Artificial Intelligence, IJCAI-97*, 474–479. Nagoya, Japan: Morgan Kaufmann.

Loiez, E. 1997. *Contribution au Diagnostic de Systèmes Analogiques*. Ph.D. Dissertation, Université des Sciences et Technologies de Lille.

Patton, R., and Chen, J. 1991. A review of parity space approaches to fault diagnosis. In *IFAC SAFEPROCESS Symposium*.

Raiman, O. 1992. The alibi principle. In Hamscher et al. (1992). 66–70.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32(1):57–96.

Staroswiecki, M., and Comtet-Varga, G. 2000. Design of structured residuals in algebraic dynamic systems. In *IFAC Safeprocess'2000*.

Struss, P. 1997. Fundamentals of model-based diagnosis of dynamic systems. In *15th International Joint Conference on Artificial Intelligence, IJCAI-97*, 480–485. Nagoya, Japan: Morgan Kaufmann.