Robust fault detection using consistency techniques with application to an automotive engine

Esteban R. Gelso ∗ , Erik Frisk∗∗ and Joaquim Armengol∗

∗ Institut d’Informàtica i Aplicacions, Universitat de Girona, Campus de Montilivi, E-17071 Girona, Spain (e-mail: {esteban.gelso,joaquim.armengol}@udg.edu).
** Dept. of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden (e-mail: frisk@isy.liu.se)

Abstract: Monitoring of the air intake system of an automotive engine is important to meet emission related legislative diagnosis requirements. In this paper, the problem of fault detection in the air intake system is stated as a constraint satisfaction problem over continuous domains with a big number of variables and constraints. This problem can be solved using Consistency Techniques. Consistency techniques are shown to be particularly efficient for checking the consistency of the Analytical Redundancy Relations (ARRs), dealing with uncertain measurements and parameters, and using experimental data.

Keywords: Fault detection; Consistency; Uncertainty; Interval analysis; Automotive engines.

1. INTRODUCTION

Automotive engines is an important application for model-based diagnosis not only because of environmentally based legislative regulations, but also because of repairability, availability, and vehicle protection (Nyberg, 2002). Different model-based approaches have been studied in several works as in (Gertler et al., 1995; Nyberg and Nielsen, 1998; Nyberg et al., 2001; Nyberg, 2002; Kimmich et al., 2005). One important part of the diagnosis requirements for automotive engines is the air path. Possible faults include sensor faults, actuator faults and leakages. These types of faults typically lead to degraded emission control, and also possible damage to engine components.

This paper introduces a fault detection method based on a model that takes into account the uncertainties in the measured signals and in the model by using intervals. These uncertainties are caused by, for example, non-modeled effects, electrical disturbances, model simplifications, and so on.

Two research communities work on model-based techniques: the FDI (Fault Detection and Isolation) community, formed by researchers with a background in control systems engineering, and the DX (Principles of Diagnosis) community, formed by researchers with a background in computer science and intelligent systems.

Among the techniques developed by the FDI research community, there are classical methods, such as state observers, parity equations and parameter estimation (Blanke et al., 2003; Chen and Patton, 1998; Gertler, 1998; Patton et al., 2000). One of the methods to detect faults consists in comparing the behavior of an actual system and a model of the system. This principle is called analytical redundancy. Consider an actual system or a part of it that can be represented by a model described by the following nonlinear discrete-time equation,

\[ y(k) = f(y(k-1), \ldots, y(k-n), u(k-1), \ldots, u(k-m), \theta), \]

where \( y(k) \in \mathbb{R}^n \) and \( u(k-n) \in \mathbb{R}^n \) are the outputs of the system at instants \( k \ldots k-n \), \( f \) is a vector of functions, \( u(k-1) \in \mathbb{R}^m \ldots u(k-m) \in \mathbb{R}^m \) are the inputs at instants \( k-1 \ldots k-m \), and \( \theta \in \mathbb{R}^p \) is a vector of parameters.

An analytical redundancy relation (ARR) is an algebraic constraint deduced from the system model which contains only measured variables. An ARR for Equation 1 is

\[ y(k) = \hat{y}(k), \]

where \( y(k) \) is the measured output of the system at instant \( k \) and \( \hat{y}(k) \) is the analytical output of the model at instant \( k \).

An ARR is used to check the consistency of the observations with respect to the system model. Therefore, a fault is detected when

\[ \hat{y}(k) \neq y(k), \]

or equivalently

\[ r(k) = \hat{y}(k) - y(k) \neq 0, \]

where \( r \) is called the residual of the ARR.

The main problem is that the measured output \( y(k) \) and the computed output \( \hat{y}(k) \) are seldom the same because the model is, by definition, inaccurate, i.e. it is an approximate representation of the system. This is the consequence of the uncertainties of the system and the procedure of systems modelling.

The better model used to represent the dynamic behavior of the system, the better will be the chance of improving the reliability and performance in detection and diagnosis.
of faults. However, modelling errors and disturbances in complex engineering systems are unavoidable and, hence there is a need to develop robust fault diagnosis algorithms. The goal of robustness is to minimize the false and missing alarm rates due to the effects that modelling uncertainty and unknown disturbances will have on the residuals. This can be achieved in several ways, e.g. by statistical data processing, averaging, or by finding and using the most effective threshold. One way to find effective thresholds is using intervals to bound the uncertainty of parameters and measurements. In this way adaptive thresholds (envelopes) could be obtained.

Some interval methods have been proposed in the context of fault detection and diagnosis, e.g. (Armengol et al., 2000; Ploix and Follot, 2001; Puig et al., 2006b). (Stancu et al., 2003) includes constraint propagation to solve fault detection problems. In (Puig et al., 2006a), the problem is solved using a tool known as IntervalPeeler, based on constraint projection algorithms (2B-consistency) to reduce interval domains of variables without bisections.

In this paper, the uncertainties associated with the system itself and with the measurements are taken into account, also by using intervals.

When interval uncertainties are considered, consistency methods which combine interval methods and constraint satisfaction techniques can be used to solve different problems such as parameter and state estimation. Constraint satisfaction techniques implement local reasoning on constraints to remove inconsistent values from variable domains. In practice, the set of inconsistent values is computed by means of interval reasoning. In section 2 the alternative to use local and global consistency techniques such as Hull and Box consistency, is explored. The fault detection problem is shown like a constraint satisfaction problem and the resolution of this problem is performed by the solver RealPaver (Granvilliers and Benhamou, 2006). A stronger technique than 2B-consistency, called Weak-3B, is used.

The novelty of this paper is that an interval observer is stated as a Constrain Satisfaction Problem (CSP), in order to solve the fault detection problem by means of consistency techniques. Also, a sliding time window is used to reduce the computational effort. Thus, the aim of this paper is to show the usefulness of the consistency methods to solve real and highly complex fault detection problems like the ones of automotive engines.

In section 3, the engine used model is described, and the experimental fault detection results in two scenarios are presented. An alternative approach using the signs of the symptoms is introduced as well. Finally, section 4 provides some conclusions and outlines the future work.

2. FAULT DETECTION AS A CONSTRAINT SATISFACTION PROBLEM

Many engineering problems can be formulated in a logical form by means of some kind of first order predicate formulas: formulas with the logical quantifiers (universal and existential), a set of real continuous functions (equalities and inequalities), and variables ranging over real interval domains.

As defined in (Shary, 2002), a numerical constraint satisfaction problem is a triple CSP = (V, D, C(x)) defined by

1. a set of numeric variables V = {x1, . . . , xn},
2. a set of domains D = {D1, . . . , Dn} where Dk, a set of numeric values, is the domain associated with the variable xk,
3. a set of constraints C(x) = {C1(x), . . . , Cm(x)} where a constraint Ck(x) is determined by a numeric relation (equation, inequality, inclusion, etc.) linking a set of variables under consideration.

The fault detection problem can be represented by a CSP similar to the one presented in (Jaulin, 2002), which deals with the problem of nonlinear state estimation. For example, the set of variables for the system (1) with one output is

\[ V = \{ \theta_1, . . . , \theta_m, y(k-n), . . . , y(k), u_1(k-m), . . . , u_1(k-1), . . . , u_n(k-m), . . . , u_n(k-1) \} \]

the set of domains is

\[ D = \{ \Theta_1, . . . , \Theta_m, Y(k-n), . . . , Y(k), U_1(k-m), . . . , U_1(k-1), . . . , U_n(k-m), . . . , U_n(k-1) \} \]

and the set of constraints is

\[ C = f(y(k-1), . . . , y(k-n), u(k-1), . . . , u(k-m), \theta) - y(k) = 0 \] Consistency techniques can be used to contract the domains of the variables involved removing inconsistent values (Collavizza et al., 1999; Benhamou et al., 1999). In particular for the fault detection application, they are used to guarantee that the observed behavior and the model are inconsistent when there is no solution. The algorithms that are based on consistency techniques are actually “branch and prune” algorithms, i.e., algorithms that can be defined as an iteration of two steps (Collavizza et al., 1999):

1. Pruning the search space by reducing the intervals associated with the variables until a given consistency property is satisfied.
2. Generating subproblems by splitting the domains of a variable

Most interval constraint solvers are based on either hull-consistency (also called 2B-consistency) or box-consistency, or a variation of them (Benhamou et al., 1999). Box-consistency tackles the problem of hull-consistency for variables with many occurrences in a constraint. The aforementioned techniques are said to be local: each reduction is applied over one domain with respect to one constraint. Better pruning of the variable domains may be achieved if complementary to a local property, some global properties are also enforced on the overall constraint set.

In this paper, the solution of the fault detection CSP is performed by using the solver RealPaver (Granvilliers and Benhamou, 2006). Weak-3B consistency, a stronger technique than hull consistency and box consistency, is used in Section 3.

2.1 Diagnostic observer

When dynamics is present and when a model’s estimates of states are improved by feedback from measured signals, it is called an observer. An observer used for diagnosis is called a diagnostic observer.
Taking into account the uncertainty by means of intervals, as defined in (Puig et al., 2006b), a non-linear interval observer equation with a Luenberger-like structure for a system in the state-space representation can be written as:

\[
\begin{align*}
\dot{\hat{x}}(k+1) &= \mathbf{g}(\hat{x}(k), \mathbf{u}(k), \theta) + \mathbf{K}(y(k) - \hat{y}(k)), \\
\hat{y}(k) &= h(\hat{x}(k), \mathbf{u}(k), \theta),
\end{align*}
\]

where \(\hat{x} \in \mathbb{R}^{nx}\) and \(\hat{y} \in \mathbb{R}^{ny}\) are estimated state and output vectors of dimension \(nx\) and \(ny\), respectively, \(\mathbf{u} \in \mathbb{R}^{nu}\) and \(y \in \mathbb{R}^{nu}\) are measured input and output vectors of dimension \(nu\) and \(ny\), \(\theta\) is the vector of uncertain parameters of dimension \(np\) with their values bounded \(\theta \in \mathbb{R}^{np}\), and \(\mathbf{K}\) is the gain of the observer. The choice of \(K\) can be done, for example, by pole placement. The observer functions like a low-pass filter and thus the pole placement is a compromise between fast fault response and sensitivity to disturbances and noise (Nyberg and Nielsen, 1998).

The estimated outputs are used to check the consistency of the observations with respect to the system model. Therefore a fault is detected when the measured value is either larger or smaller than the predicted value or in other words, when the output of the model is not consistent with the measured output. This assertion is expressed through the logical statement:

\[
(\forall y(k) \in Y(k)) (\forall \hat{y}(k) \in \hat{Y}(k)) \mathbf{r}(k) \neq 0,
\]

where \(\mathbf{r}(k) = \hat{y}(k) - \hat{y}(k)\) is a vector of residuals.

The dynamic system (5) can be represented as a CSP:

\[
\begin{align*}
\mathcal{V} &= \{\theta, y(1), \ldots, y(k), \hat{y}(1), \ldots, \hat{y}(k), \hat{x}(1), \ldots, \hat{x}(k+1), \mathbf{u}(1), \ldots, \mathbf{u}(k)\} \\
\mathcal{D} &= \{\mathbf{Y}(0), \ldots, \hat{Y}(0), \ldots, \hat{Y}(k), \hat{X}(0), \ldots, \hat{X}(k+1), \mathbf{U}(0), \ldots, \mathbf{U}(k)\} \\
\mathcal{C} &= \{\mathbf{g}(2) = \mathbf{g}(\hat{x}(1), \mathbf{u}(1), \theta) + \mathbf{K}(y(1) - \hat{y}(1)) \\
&\quad \mathbf{h}(1) = h(\hat{x}(1), \mathbf{u}(1), \theta) \\
&\vdots \\
&\hat{x}(k+1) = g(\hat{x}(k), \mathbf{u}(k), \theta) + \mathbf{K}(y(k) - \hat{y}(k)) \\
&\hat{y}(k) = h(\hat{x}(k), \mathbf{u}(k), \theta)\}
\end{align*}
\]

A problem finding the CSP solution is the continuous increment with time in the computational effort. As it is applied in this paper (Section 3.3), an alternative to overcome this problem is the use of a sliding time window. The time interval from the initial time point to the current one is called time window of length \(w\).

3. APPLICATION TO THE AIR INTAKE SYSTEM OF AN AUTOMOTIVE ENGINE

3.1 System description

A schematic picture of the air-intake system is shown in Fig. 1. Ambient air enters the system and an air-mass flow sensor measures the air-mass flow rate \(W_a\). Next, the air passes the compressor side of the turbo-charger, the intercooler and then the throttle. The flow \(W_{2b}\) is dependent on the intercooler and manifold pressures, \(p_{ic}\) and \(p_{im}\), the temperature \(T_{ic}\), and the throttle angle \(\alpha\).

Finally the air enters the cylinder and this flow, \(W_{eq}\), is dependent on \(p_{em}\) and \(p_{cem}\), the temperature \(T_{im}\), the engine speed \(N\) and the air-fuel ratio \(\lambda\).

![Fig. 1. A schematic figure of the turbo-charged engine including the sensors.](image)

The sensor signals, that are available to the diagnosis system, are listed in Table 1.

<table>
<thead>
<tr>
<th>Pressure sensors</th>
<th>Temperature sensors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Symbol</td>
</tr>
<tr>
<td>Ambient pressure</td>
<td>(p_a)</td>
</tr>
<tr>
<td>Before compressor</td>
<td>(p_{af})</td>
</tr>
<tr>
<td>After compressor</td>
<td>(p_{af})</td>
</tr>
<tr>
<td>After intercooler</td>
<td>(p_{ic})</td>
</tr>
<tr>
<td>Intake manifold</td>
<td>(p_{in})</td>
</tr>
<tr>
<td>Exhaust manifold</td>
<td>(p_{en})</td>
</tr>
<tr>
<td>After Turbine</td>
<td>(p_t)</td>
</tr>
<tr>
<td>Air-mass flow</td>
<td>(W_{af})</td>
</tr>
<tr>
<td>Throttle angle</td>
<td>(\alpha)</td>
</tr>
<tr>
<td>Engine speed</td>
<td>(N)</td>
</tr>
<tr>
<td>Turbine speed</td>
<td>(N_t)</td>
</tr>
</tbody>
</table>

Table 1. Available sensors signals.

The faults in the air-intake system can be, for instance, boost leakage, manifold leakage, pressure sensor bias, pressure sensor gain-fault, etc, as described in (Nyberg, 2002).

3.2 Model equations

The model used in this paper is a part of the Mean Value Engine Model explained in (Andersson, 2005). This model describes the average behavior of the engine over one to several thousands of engine cycles, and is a component based model in which each component is described in terms of equations, constants, parameters, states, inputs and outputs. The equations describing the fault free air intake model can be written as
\[ \frac{dp_{\text{im}}}{dt} = \frac{R_e T_{\text{im}}}{V_{\text{im}}} (W_{\text{th}} - W_{\text{cyl}}) + \frac{m_{\text{im}} R_a}{V_{\text{im}}} \frac{dT_{\text{im}}}{dt} \]  
(7)

\[ W_{\text{th}} = \frac{p_{\text{ic}}}{\sqrt[\gamma - 1]{\Pi}} \Psi(\Pi) A_{\text{eff}}(\alpha) \]  
(8)

\[ W_{\text{cyl}} = p_{\text{im}} C_1 \left( 1 + \frac{r_c - \left( \frac{p_{\text{im}}}{p_{\text{ic}}} \right)^{1/\gamma}}{r_c - 1} \right) \left( V_d N \right) \frac{R_{\text{im}}}{T_{\text{im}}} \]  
(9)

where

\[ \Pi = \frac{p_{\text{im}}}{p_{\text{ic}}} \]  
(10)

\[ \Psi^*(\Pi) = \sqrt{\frac{2\gamma}{\gamma - 1} (\Pi^2 - \Pi + 1)} \]  
(11)

\[ \Psi(\Pi) = \begin{cases} 0 & \text{if } \Pi \leq \left( \frac{2}{\gamma + 1} \right)^{1/\gamma} \\ \Psi^*(\Pi) & \text{if } \left( \frac{2}{\gamma + 1} \right)^{1/\gamma} \leq \Pi_{\text{lin}} \\ \Psi^*(\Pi_{\text{lin}})(\Pi - 1) & \text{if } \Pi_{\text{lin}} < \Pi \leq 1 \end{cases} \]  
(12)

The interval method presented in this paper uses discrete-time models, in this case a discretization is obtained by using a first order approximation:

\[ x(t + T_s) \simeq x(t) + T_s g(x(t), u(t), \theta), \]  
(13)

where, from (7) and including a non-linear interval observer, it is obtained:

\[ \hat{p}_{\text{im}}(k + 1) = \hat{p}_{\text{im}}(k) + \frac{R_e T_{\text{im}}(k)}{V_{\text{im}}} (\hat{W}_{\text{th}}(k) - \hat{W}_{\text{cyl}}(k)) \\ + K (p_{\text{im}}(k) - \hat{p}_{\text{im}}(k)) \]  
(14)

where the set of sensors considered are: pressures \( p_{\text{im}}, p_{\text{ic}} \) and \( p_{\text{em}} \), temperatures \( T_{\text{im}} \) and \( T_{\text{ic}} \), engine speed \( N \) and throttle plate angle \( \alpha \).

The uncertain parameters selected are two engine specific parameters, and those are the gain parameter \( C_1 \), which describes the engine pumping capabilities, and the ratio of specific heats \( \gamma \). They have been bounded using the criterion that in the fault free case, there should be no false alarm. The variable \( \lambda \) (the air-fuel ratio) has been considered as an interval, instead of the measured value, because of the accuracy of the sensor and for a sake of simplicity.

The set of variables of this model represented as a CSP is

\[ \mathcal{V} = \{ C_1, \gamma, \lambda(k - w), \ldots, \lambda(k - 1), \hat{p}_{\text{im}}(k - w), \ldots, \hat{p}_{\text{im}}(k), p_{\text{im}}(k - w), \ldots, p_{\text{im}}(k - 1), p_{\text{ic}}(k - w), \ldots, p_{\text{ic}}(k - 1), p_{\text{em}}(k - w), \ldots, p_{\text{em}}(k - 1), T_{\text{ic}}(k - w), \ldots, T_{\text{ic}}(k - 1), T_{\text{im}}(k - w), \ldots, T_{\text{im}}(k - 1), N(k - w), \ldots, N(k - 1), \alpha(k - w), \ldots, \alpha(k - 1) \}, \]

and the set of initial domains for the estimated variable \( \hat{p}_{\text{im}} \) has been taken equal to \([1 \times 10^4, 2 \times 10^5]\) with the exception of the initial domains of \( \hat{p}_{\text{im}}(k - w) \) and \( \hat{p}_{\text{im}}(k) \), at the beginning and the end of the time window, which have been assigned a value equal to the interval measurements \( p_{\text{im}}(k - w) \) and \( p_{\text{im}}(k) \).

### 3.3 Experimental results

All experiments were performed on a four-cylinder turbocharged spark-ignited SAAB engine located in the research laboratory at Vehicular Systems Group, Linköping University. The engine is mounted in a test bench together with a Schenck dynamometer.

In this section two faulty scenarios are considered, (i) a gain-fault in the sensor of pressure \( p_{\text{ic}} \), and (ii), a gain-fault in the engine speed sensor. The fault detection results are obtained by using Weak-3B consistency technique and a window length equal to 30 samples (0.3s). The computation time required and the sample time have the same order of magnitude.

When no solution is found to the CSP, a fault is detected. Otherwise, when the observed behavior and the model are not proven to be inconsistent, means there is not a fault or it could not be detected. In this way, the proposed approach prioritizes to avoid false alarms to missed alarms.

#### • First scenario

In Fig. 2, obtained results in the case of no fault and a 10% gain-fault in the pressure sensor of \( p_{\text{ic}} \) are shown. A “1” indicates there is a fault and a “0” means there is not a fault or it could not be detected. As shown in this figure, there is no false alarm in absence of fault. The fault in the sensor begins at sample 600 and is detected at sample 604.

![Fig. 2. First scenario fault detection. Top: no fault. Bottom: gain-fault in the sensor of pressure \( p_{\text{ic}} \) beginning at sample 600. The fault is detected from sample 604.](image-url)
than the sample time being not suitable to operate in real-time; it can be used when a fault is detected to obtain more information, and then, to improve the task of diagnosis (Section 3.4).

• Second scenario
Fig. 4 shows the results in the case of no fault and a 10% gain-fault in the pressure sensor of $p_{im}$. The fault in the sensor begins at sample 800 and is detected at the same time as the fault.

Fig. 5 shows the interval measurement and the estimated manifold pressure in the fault free situation of this scenario.

3.4 Diagnosis: signs of the symptoms

When it is possible to utilize detailed models for the faults, this information can be used together with the signs in the residuals, to prune the candidate space when performing the fault diagnosis task, as proposed in (Calderón-Espinoza et al., 2007).

This approach could be applied to perform the diagnosis in both studied scenarios. In order to do this, it is needed to:

• Include in the fault signature matrix, the influence of the faults in the residuals, and

• Obtain the sign of the symptom. This could be obtained by observing the behavior of the estimated output with respect to the measurement. For instance, the sign would be +1 if the estimation is greater than the interval measurement, or if the estimation is smaller than the interval measurement, the sign would be -1.

For both scenarios, when a fault is detected, the algorithm estimates the manifold pressure at the end of each sliding window and the consistent region of this variable can be seen in Fig 6 and 7. As it is expected, the interval measurement (solid line) does not intersect with the estimate (dashed line), and for the first case, the estimates are always smaller than the measurements, whereas for the second case, the opposite relation is observed.

4. CONCLUSIONS AND FUTURE WORK

When interval uncertainties are considered, consistency methods can be used to solve fault detection problems. In this paper, consistency methods are used to increase robustness of a diagnosis system for an automotive engine application. In this paper through the obtained results, consistency techniques are shown to be particularly efficient to check the consistency of the Analytical Redundancy Relations (ARRs) and diagnostic observers, dealing with uncertain measurements and parameters. In the future, diagnosis approach introduced in section 3.4, which uses the signs of the symptoms, must be studied in depth in order to perform this task.

ACKNOWLEDGEMENTS

This work has been funded by the Spanish Government (Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, Ministerio de Educación y Ciencia) through the coordinated research project grant No. DPI2006-15476-C02-02, by the grant No. 2005SGR00296

Fig. 3. First scenario without faults. The upper plot shows measured and estimated manifold pressure.

Fig. 4. Second scenario fault detection. Top: no fault. Bottom: gain-fault in the sensor of pressure $p_{im}$. The fault is detected at the same time as the fault.

Fig. 5. Second scenario without faults. The upper plot shows measured and estimated manifold pressure.
Fig. 6. First scenario with a gain-fault in the sensor of pressure $p_{ic}$. Starts at 600, detected at 604.

and the Department d’Innovació, Universitats, i Empresa of the Government of Catalonia.

REFERENCES


