Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Simulation based Evaluation of Fault Detection Algorithms with Applications to Wear Diagnosis in Manipulators ⋆

Andreas Samuelsson [***] **André Carvalho Bittencourt** [*]
Kari Saarinen [***] **Shiva Sander-Tavallaey** [***]
Mikael Norrlöf [*,**] **Hans Andersson** [**] **Svante Gunnarsson** [*]

[*] *Department of Electrical Engineering, Linköping University, Sweden*
andrecb@isy.liu.se
[**] *ABB Robotics, Västerås, Sweden*
[***] *ABB Corporate Research, Västerås, Sweden*

Abstract: Fault detection algorithms (FDAs) process data to generate a test quantity. Test quantities are used to determine presence of a fault in a monitored system, despite disturbances. Because only limited knowledge of the system can be embedded in an FDA, it is important to evaluate it in scenarios relevant in practice. In this paper, simulation based approaches are proposed in an attempt to determine: *i*) which disturbances affect the output of an FDA the most; *ii*) how to compare the performance of different FDAs; and *iii*) which combinations of fault change size and disturbances variations are allowed to achieve satisfactory performance. The ideas presented are inspired by the literature of design of experiments, surrogate models, sensitivity analysis and change detection. The approaches are illustrated for the problem of wear diagnosis in manipulators where three FDAs are considered. The application study reveals that disturbances caused by variations in temperature and payload mass error affect the FDAs the most. It is also shown how the size of these disturbances delimit the capacity of an FDA to relate to wear changes. Further comparison of the FDAs reveal which performs "best" in average.

Keywords: Fault detection and diagnosis; Sensitivity analysis; Robotics.

1. INTRODUCTION

Fault detection and fault diagnosis can be used to improve safety, reliability, availability, and maintainability of technical systems [Isermann, 2006]. In fault detection, observations from the system, e.g. data, are processed and compared to available knowledge of the system to generate symptoms. Symptoms are a partial diagnose of the system, i.e. a statement about which states of the system could possibly explain the current observations. The diagnosis of complex systems typically makes use of several fault detection methods, each containing partial information of the system. In fault diagnosis, the different symptoms are processed to generate a statement of the state (condition) consistent to all observations and knowledge embedded in the diagnosis solution.

While increasing the amount of symptoms used for fault diagnosis may increase the quality of the diagnostic process, it is clear that the accuracy of the symptoms are crucial. The design and verification of fault detection methods are therefore important. Fig. 1, shows an overall scheme of a fault detection scheme. The monitored system is affected by input factors which are relevant for diagnostics, e.g. faults and disturbances, and generates observations. The observations are processed to extract relevant features that can describe the status of the system (e.g. parameters, residuals, signal spectra). The behavior of the features are then compared to (known) reference behaviors (e.g. based on distances) to generate a test quantity. The combined tasks of feature extraction and behavior comparison is denoted fault detection algorithm (FDA). Finally, a decision rule (e.g. a threshold check or a statistical test) is used to accept or reject the reference behaviors that the test quantity can explain, i.e. it generates a symptom.

⋆ This work was supported by ABB and the Vinnova Industry Excellence Center LINK-SIC at Linköping University.

1.1 Problem Description and Motivation

The accuracy of the symptoms generated by fault detection is determined by the ability of the test quantity generated by the FDA to relate to changes in the system behavior. It is thus natural to evaluate fault detection methods based on the test quantities alone, independent of the decision rule used. This is for instance in line with the theory of statistical hypothesis testing, when an optimal test is determined only by the statistical behavior of the test quantity, see e.g. Van Trees and Kristine [2013].

It is considered that a test quantity, denoted y, is a *scalar* that measures deviations from one or more reference behaviors. The reference behaviors can be represented by states of interest, e.g. different fault types. In this work, the focus is on the analysis of a single fault, denoted f. Rather than considering test quantities which are time sequences, e.g. generated from residuals, the focus is restricted to *batch fault detection algorithms*, which produce a scalar y

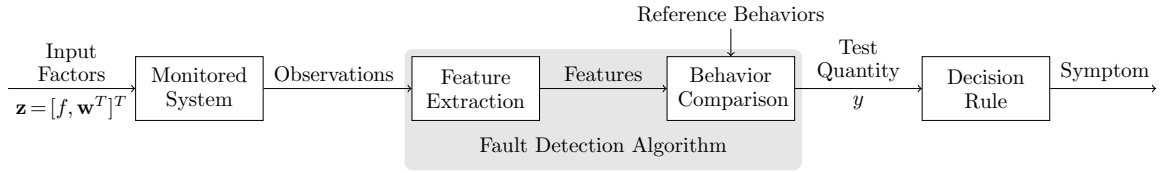


Fig. 1. Overview of a fault detection scheme. The monitored system is affected by input factors and generates observations. Features are extracted from the observations which are compared against reference (known) behaviors of the features to generate test quantities. A decision rule determines which behaviors better explain the observations, i.e. it generates a symptom.

for an entire data batch. Batch methods are common for signal/data-driven approaches and parameter estimation, but similar ideas could be used also for time sequences by summarizing the sequence to a scalar, e.g. by considering steady-state values or some norm.

In practice, the data input to fault detection (and thus y) are not only affected by f but by a **collection of n factors** $\mathbf{z} = [f, \mathbf{w}^T]^T$, where $\mathbf{w} = [w_1, \cdots, w_i, \cdots, w_{n-1}]^T$ relates to nuisance factors, e.g. disturbances. The nuisance factors, $\mathbf{w}$, may cause undesired variations in y, deteriorating its capacity to distinguish changes in f and thus complicating a decision. Under specified conditions and assumptions, optimality of FDAs might be possible, see e.g. Frank and Ding [1997], Li and Zhou [2009], Liu and Zhou [2008], Wei and Verhaegen [2011], and it may be possible to compare different schemes [Isermann, 1994, 2006]. However, since only partial knowledge of the system can be embedded in any FDA, it is important to evaluate it in scenarios which are relevant for its practical use. From a practical perspective, given a complex system and candidate FDAs, the following questions are of relevance:

Q-1 Which factors in $\mathbf{w}$ affect y the most? And should therefore be given more relevance for further development of the FDA.
Q-2 How can test quantities generated from different FDAs be compared and evaluated against each other to enable selection of the "best" FDAs?
Q-3 What is the effective scope of an FDA? That is, for what region in the $\mathbf{z}$ space is the ability of y to relate to f satisfactory?

Notice that the focus is not on properties of a particular FDA but to define approaches to evaluate and compare any FDA.

These questions can be addressed at different levels of closeness to the real application. Level 0 corresponds to the ideal case where the FDAs are evaluated with operational data. This is particularly difficult since it may take extremely long times for faults to appear. To overcome this, data can be collected from experiments performed in a lab, where faults and disturbances are induced, corresponding to Level 1 studies. Even at Level 1, an extensive evaluation is often inviable due to the extreme costs and time required. Furthermore, it is often the case that all (or parts of) the factors $\mathbf{z}$ are unmeasurable and therefore a complete analysis based on real data is difficult. At Level 2, data are generated based on a *simulations* of the monitored system, which is a more viable alternative. The simulation study must, on the other hand, be designed carefully so that it is representative of scenarios of practical relevance.

1.2 Main Contributions and Outline

In this paper, ideas inspired by the literature of *design of experiments*, *sensitivity analysis* and *change detection* are presented to address these questions based on simulation studies. Even in simulation studies, an extensive analysis of the effects of $\mathbf{z}$ to y may exhaust the computational resources and time available. An important idea considered here is to *bypass the need for simulation/experimental data* using a surrogate (or meta) model. Different types of surrogate models are possible, e.g. based on neural networks and Gaussian processes. For its simplicity and tractability, the **surrogate models** considered here will take the form of a linear regression,

$$y = \varphi(\mathbf{z})^T \boldsymbol{\theta} + \varepsilon, \quad \varphi(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^{n_\theta}, \quad y, \varepsilon \in \mathbb{R} \qquad (1)$$

where the regressors function $\varphi(\cdot)$ makes a direct map from $\mathbf{z}$ to y through the regression coefficients $\boldsymbol{\theta}$ and ε is an additive uncertainty term. The surrogate model incorporates both the monitored system and FDA. Studies based on surrogate models are denoted as Level 3. In such approach, the choice of factors $\mathbf{z}$ and regressors, the identification of $\boldsymbol{\theta}$ and model validation are important and are subject of study in the field known as *design of experiments* (DOE) which is presented in Sec. 2.

An answer to Q-1 is presented in Sec. 3, where the coefficients $\boldsymbol{\theta}$ of the regression models are studied using sensitivity analysis to determine which factors in $\mathbf{w}$ affect y the most. A main advantage with the use of surrogate models is that Monte Carlo (MC) simulations can be performed efficiently. MC runs are used in Sec. 4 to evaluate a measure of average effects of changes in y by f which is used to address Q-2. In Sec. 5, a measure of satisfactory performance is suggested which is evaluated with MC runs under various combinations of $\mathbf{z}$ in an attempt to answer Q-3. In Sec. 6, the ideas are illustrated for the evaluation of methods used for wear diagnosis in industrial robots. Relevant characteristics of the problem and methods are revealed from the study. Concluding remarks are given in Sec. 7.

2. DESIGN OF EXPERIMENTS

Design of experiments (DOE) can be applied to any system where the experimenter has control over the input variables, or *input factors*, and that the output can be measured [Kleijnen et al., 2005]. Here, the object of study is an FDA applied in combination to a monitored system. The input factors are, e.g., faults and disturbances present in the monitored system, and the output is the test quantity y. The next sections are organized to give an introduction to the field, for more details see e.g. Box et al. [1978], Kleijnen et al. [2005], Sanchez [2006].

2.1 Choice of Input Factors

The first task in designing an experiment is the selection of the input factors $\mathbf{z}$ and their possible range of values. Factors can be included according to the objectives of the study to verify or falsify assumptions about the behavior of the test quantity and to study their relations in detail. The choice of factors should be performed carefully, with the help of experts in the application, since a poor specification may generate misleading results. For the context here, input factors include the fault f and nuisance factors $\mathbf{w}$ such as external disturbances, operating points, etc.

Once the factors are chosen, the experimenter must decide their range of values and a discrete set of *factor levels* that shall be considered in the study. A more detailed study is possible by increasing the **number of levels**, m, in a compromise with the number of experiments required. The factor levels chosen will have an impact on the study and it is therefore very important to choose levels which are extreme but not impossible for realistic situations. Two representations of factor levels are typically used:

- **natural levels** are the values for the factors that are used in the experiment or simulation;
- **coded levels** all factors are normalized to the same scale. Used when *identifying* the surrogate models.

2.2 Surrogate Models as Linear Regressions

Linear regression models as in (1) are simple, tractable and easy to interpret. For these reasons, they are a popular choice in the DOE literature. A limitation is that they may misrepresent the relations between $\mathbf{z}$ and y. To circumvent this, more complex model structures, such as neural networks and Gaussian processes could be considered, see e.g. [Oakley and O'Hagan, 2004]. Compared to linear regressions, more complex models may be less interpretable and tractable which are important characteristics for surrogate models. Easy to interpret models are particularly important for sensitivity analysis, studied in Sec. 3.

Many different model structures of linear regressions can be considered. From the DOE literature, two structures are commonly used. A **main effects model** has regressors that are directly dependent on the inputs factors, i.e.

$$\varphi(\mathbf{z})^T = [\,1\ \mathbf{z}^T\,] = [\,1\ f\ w_1\ \cdots\ w_{n-1}\,] \tag{2a}$$

$$\boldsymbol{\theta}^T = [\,b\ \{\theta_i\}\,] \tag{2b}$$

where b is a bias term and θ_i has indices $i \in [0, \cdots, n-1]$. Since this is a simple model it may not be a realistic representation of the system. A **second-order model** extends the main-effects model with *interaction* (cross) and *quadratic* terms as

$$\varphi(\mathbf{z})^T = [\,1\ \mathbf{z}^T\ \mathrm{svec}(\mathbf{z}\mathbf{z}^T)^T\,] \tag{3a}$$

$$\boldsymbol{\theta}^T = [\,b\ \{\theta_i\}\ \{\theta_{ij}\}\ \{\theta_{ii}\}\,] \tag{3b}$$

where $\mathrm{svec}(\cdot)$ maps a symmetric matrix of size N to a vector of length $N(N+1)/2$ and $i, j \in [0, \cdots n\text{–}1]$ with $i > j$. A second order model can capture more complex relations between the factors than a main effects model. However, since each factor is included in several terms, it is more difficult to analyze the effects of different factors to y. Notice that the models can be extended further with three way interactions and terms of higher orders.

2.3 Identification

Consider that N experiments are performed with inputs

$$Z = [\mathbf{z}_1, \cdots, \mathbf{z}_N]^T \in \mathbb{R}^{N \times n} \tag{4}$$

and outputs $\mathbf{y} = [y_1, \cdots, y_N]^T \in \mathbb{R}^N$. Given that the test quantities can be described by (1), the resulting model is $\mathbf{y} = \Phi(Z)\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with $\Phi(Z) \triangleq [\varphi(\mathbf{z}_1), \cdots, \varphi(\mathbf{z}_N)]^T$. To find the coefficients $\boldsymbol{\theta}$, a least-squares error criterion gives

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \Phi(Z)\,\boldsymbol{\theta}\|_2^2 \\ &= [\Phi(Z)^T\,\Phi(Z)]^{-1}\Phi(Z)^T\mathbf{y}.\end{aligned} \tag{5}$$

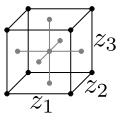
Notice again that the **coded levels should be used** when identifying the regression coefficients. Otherwise, the scaling of the variables will hinder some of the analyses presented further.

2.4 Design Matrix

A design matrix represents the user choice of simulation experiments to be performed. Typically, the columns correspond to the factor levels and rows are *design points*, i.e. a specific choice of the coded levels $\mathbf{z}$. Using the previously introduced notation, a design matrix corresponds to a specific choice of Z in (4).

A *full factorial design* considers all possible combinations of the m levels and n factors possible, giving m^n experiments. A *fractional factorial design* considers only a fraction m^{n-p} of the full factorial design, thus reducing the number of experiments by m^p. A **central composite design** extends a two level, i.e. $m = 2$, factorial design (full or fractional), with points at the center of the factor levels and $2n$ "star points" which represents extreme values for the factors, see the example. A central composite design allows for estimation of higher order models with small number of experiments. For more, see e.g. Atkinson et al. [2007], Fedorov [1972], Kleijnen et al. [2005].

Example. A $n = 3$ central composite design based on 2^3 full factorial design (black), a center point and star points at the faces (gray).



2.5 Design Parameters

The validity of a surrogate model is of course limited. For example, it should not be expected that the same model can be used for different monitored systems or for different FDAs. The settings that determine the validity of the surrogate models are called *design parameters*. One surrogate model should be identified for each different combination of design parameters.

2.6 Model Validation

The analyses performed in this paper are based on surrogate models and it is therefore important to validate them. Model validation is used to assess whether the model will generalize to input values independent of those used during the model identification. Model validation can be done by *cross-validation*, where a fresh dataset, denoted Z_v, is used with the sole purpose of validation.

The **model fit** [Ljung, 1998] can be used as a criterion to assess the validity of a model. It is defined as

$$\text{fit} = 100\left(1 - \frac{\|\mathbf{y} - \widehat{\mathbf{y}}\|_2}{\|\mathbf{y} - \bar{y}\|_2}\right), \quad \bar{y} \triangleq \frac{1}{N}\sum_{i}^{N} y_i \tag{6}$$

where the model output $\widehat{\mathbf{y}}$ is evaluated at Z_{v}. For a linear regression $\widehat{\mathbf{y}} \triangleq \Phi(Z_{\text{v}})\widehat{\boldsymbol{\theta}}$. The model fit relates to how well the model predicts the output in average.

3. DETERMINING RELEVANT FACTORS

An approach to address Q-1 is to study how changes in a factor affect the output y. The *partial derivatives* of the surrogate model with respect to the factors $\mathbf{z}$ reveal how the first order properties of y are affected by $\mathbf{z}$. This type of study is part of **sensitivity analysis** [Saltelli et al., 2008]. For the main-effects model (2), the derivatives are given directly by the coefficients $\boldsymbol{\theta}$. Because coded levels of $\mathbf{z}$ are used in the regression models, the size of a coefficient θ_i relates to how y is affected by the associated factor. For more complex models, such as the second-order model (3), the partial derivatives depend not only on $\boldsymbol{\theta}$ but also on the values of $\mathbf{z}$ where they are evaluated. Therefore, a direct comparison of the coefficients does not have the same character as for a main effects model, but can still be used to provide insights about the behavior of y.

3.1 Normalization of Coefficients

For regressors written as

$$\varphi(\mathbf{z})^T = [\,1 \;\; f \;\; \cdots\,] \tag{7a}$$

$$\boldsymbol{\theta}^T = [\,b \;\; \theta_0 \;\; \cdots\,] \tag{7b}$$

the coefficient θ_0 relates to the *direct effect* of the fault f. To facilitate the study and comparison of coefficients, the identified coefficient vector $\widehat{\boldsymbol{\theta}}$ can be normalized as

$$\overline{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}/\widehat{\theta}_0. \tag{8}$$

In this manner, the coefficients have values relative to the direct effect of f. A normalized coefficient with $|\bar{\theta}_i| < 1$ would thus mean that f has a direct effect to y which is larger than that caused by the regressor associated with $\bar{\theta}_i$. The situation where $|\bar{\theta}_i| > 1$ is possible but undesirable (unless $\bar{\theta}_i$ also relates to f). Notice that $\bar{\theta}_0 = 1$.

3.2 Group Analysis

The normalized coefficients in (8) can be **grouped** together over a subset of the design parameters to investigate different aspects of the problem. For example, consider a problem with two design parameters corresponding to the FDA used and the system monitored. Groups formed for each FDA over all monitored systems would allow for an overall comparison of the FDAs sensitivity. On the other hand, groups formed for each monitored system could be used to reveal which systems are more difficult to perform fault detection, independent of the FDA chosen.

Suppose there are K groups, where each kth group has N_k regression models. The following matrix can be formed for the kth group

$$B^k = \left[\overline{\boldsymbol{\theta}}^1, \; \cdots, \; \overline{\boldsymbol{\theta}}^{N_k}\right]^T \in \mathbb{R}^{N_k \times n_\theta}. \tag{9}$$

Each group can be analyzed using box plots for each column (each group of coefficients) of B^k. This type of analysis is illustrated further in Sec. 6.2.

4. COMPARING FAULT DETECTION ALGORITHMS

A simple approach to address Q-2 is to analyze the *average effects* a change in f gives to y when random changes of the nuisance factors $\mathbf{w}$ are present. To proceed, a change is defined in terms of hypotheses in Sec. 4.1 and a measure of average change in the output is defined in Sec. 4.2.

4.1 Two Hypotheses

The performance of a test quantity is associated to how well it can be used to relate the presence of a change from nominal in f, irrespective of variations in the disturbances $\mathbf{w}$. Given an observation y, two hypotheses are considered. The *null hypothesis*, $\mathcal{H}_0$, represents the case where y was collected when f was nominal and the *alternative hypothesis*, $\mathcal{H}_1$, states that an abnormal change in f was present. These hypotheses can be described by the particular choices of input factors

$$\mathcal{H}_0: \quad f = f^0, \qquad \mathbf{w} \sim p(\mathbf{w}), \tag{10a}$$

$$\mathcal{H}_1: \quad f = f^0 + \Delta, \qquad \mathbf{w} \sim p(\mathbf{w}), \tag{10b}$$

where f^0 is the nominal value of f, Δ is the fault change size and $p(\mathbf{w})$ is a distribution for the (considered random) nuisance factors $\mathbf{w}$. Output values collected under the different hypotheses are denoted as $y|\mathcal{H}_0$ and $y|\mathcal{H}_1$.

4.2 A Measure of Average Effects

Denoting $[\mu^0, \mu^1]$ and $[\sigma^0, \sigma^1]$ the mean and standard deviation of $y|\mathcal{H}_0$ and $y|\mathcal{H}_1$ for the hypotheses given in (10), the **signal to noise ratio** (SNR), defined as

$$\text{SNR} \triangleq \frac{\mu^1 - \mu^0}{\sigma^1}, \tag{11}$$

relates to the average effects a change of size Δ in f causes to the test quantity in relation to effects of random variations in $\mathbf{w}$. The larger the SNR value, the easier it will be to distinguish the change in f. In order to find the quantities used in the computation of the SNR, Monte Carlo runs can be performed for different realizations of $\mathbf{w}$ until enough samples of $y|\mathcal{H}_0$ and $y|\mathcal{H}_1$ are collected for the estimation of μ^0, μ^1, σ^1. Here, the use of **regression models instead of experiments** allows for *efficient MC runs*, and the quantities can be found accurately and in short time.

4.3 Group Analysis

In a similar manner as discussed in Sec. 3.2, the SNRs can be grouped over subsets of design parameters to asses different aspects of the problem. Notice that the SNRs are already normalized quantities. The use of SNRs for comparison of test quantities is illustrated in Sec. 6.3.

5. DETERMINING THE EFFECTIVE SCOPE

To address Q-3, a measure of *satisfactory performance* of a test quantity must be defined. Once the performance criterion is defined, it is possible to investigate what region in the $\mathbf{z}$ space is the criterion fulfilled. That is, the effective scope of the test quantity can be found.

5.1 A Measure of Satisfactory Performance

For fault detection, the behavior of the test quantity should allow for a correct generation of symptoms. For a given decision rule, the accuracy of the fault detection can be defined in terms of the probabilities of false, P_f, and correct detection, P_d, of a fault presence. A natural performance criterion is thus defined according to acceptable levels of P_d and P_f. This can be tested with the function

$$\text{pass} = \begin{cases} 1, & \text{if } P_f \leq P_f' \text{ and } P_d \geq P_d', \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where P_f' and P_d' are the chosen performance requirements. For a satisfactory performance of the test quantity, low P_f and high P_d are typically desirable.

The presence of an abnormal fault can be modeled with two hypotheses, e.g. given in (10). In this case, P_f relates to accepting $\mathcal{H}_1$ when $\mathcal{H}_0$ is true, and P_d relates to accepting $\mathcal{H}_1$ when $\mathcal{H}_1$ is true. The probabilities P_f and P_d are however *dependent* on the decision rule used. Different decision rules are possible, see e.g. Gustafsson [2000]. Here, a threshold check is considered since it is one of simplest and is also a common choice. It is defined as

$$\{\text{Choose } \mathcal{H}_1 \text{ if: } y \geq \hbar. \text{ Otherwise, choose } \mathcal{H}_0\} \tag{13}$$

where $\hbar$ is a threshold. The decision rule chosen and hypotheses (10) define a **binary hypothesis test** [Van Trees and Kristine, 2013]. Let $p(y|\mathcal{H}_0)$ and $p(y|\mathcal{H}_1)$ denote the probability densities of y under the different hypotheses. For the threshold check (13) with threshold value $\hbar$, P_f and P_d can be computed as

$$P_f = \int_{\hbar}^{\infty} p(y|\mathcal{H}_0)\,\mathrm{d}y, \quad P_d = \int_{\hbar}^{\infty} p(y|\mathcal{H}_1)\,\mathrm{d}y. \tag{14}$$

Notice that according to (14), for a fixed P_f there is an associated $\hbar$ and thus a P_d. The criterion (12) can therefore be verified by first finding $\hbar$ for the limiting value P_f', computing the associated P_d and checking whether it is larger than P_d'. The hypotheses densities can be estimated given a large number of observations for $y|\mathcal{H}_0$ and $y|\mathcal{H}_1$, which can be achieved efficiently with MC runs using the regression models. Here, a kernel density estimator is used, see e.g. [Bowman and Azzalini, 1997].

5.2 Finding the Effective Scope

To find the scope of a test quantity, criterion (12) can be verified for multiple binary hypothesis tests where the hypotheses in (10) are varied. In order to simplify the analyses, one nuisance factor is varied randomly at a time, while the others are kept constant. This setup can be described by the hypotheses

$$\mathcal{H}_0 : f = f^0, \qquad w_{j\neq i} = w_j', \qquad w_i \sim p(w) \tag{15a}$$

$$\mathcal{H}_1 : f = f^0 + \Delta, \qquad w_{j\neq i} = w_j', \qquad w_i \sim p(w) \tag{15b}$$

i.e. the ith nuisance factor is varied randomly while the remaining are kept constant. By checking (12) for different values of Δ and $p(w)$ in (15), it is possible to gather understanding of the effective scope of the test quantity.

With this purpose, it might be useful to restrict how the distribution $p(w)$ can be varied. Consider for instance that $p(w)$ has zero mean and variance σ^2. By varying σ, it is then possible to study how much variability of w_i is allowed for a satisfactory performance. Considering that Δ and σ can be chosen from the discrete sets

$$\boldsymbol{\Delta} = [\Delta_1, \ldots, \Delta_{N_\Delta}]^T, \quad \boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_{N_w}]^T, \tag{16}$$

all possible combinations of (Δ, σ) define a grid of size $N_\Delta \times N_w$. The result of the performance test (12) for each pair (Δ, σ) in the grid can be stored in a binary matrix of the same size, denoted **scope matrix**. Because each entry in a scope matrix relates to whether the performance criterion is achieved, its inspection allows for a straightforward analysis of the scope of a test quantity.

5.3 Group Analysis

Scope matrices can be found for each regression model. In a similar manner as discussed in Secs. 3.2 and 4.3, scope matrices can be grouped over subsets of design parameters. Because each entry in the matrices is either zero or one, the information in the group can be summarized by summing over its scope matrices. In this case, the entry values of the resulting **group scope matrix** will correspond to how many times has successful performance been achieved for the corresponding combination of (Δ, σ) over the design parameters in the group. This type of analysis is illustrated in Sec. 6.4 for the robotics application.

6. EVALUATION OF FAULT DETECTION ALGORITHMS FOR WEAR DIAGNOSIS IN ROBOTS

The framework is illustrated for the problem of wear diagnosis in an industrial robot joint. As empirically shown in Bittencourt et al. [2011] from accelerated wear tests (Level 1 studies), wear in a robot joint can lead to variations of friction. Since the friction torques must be overcome by the motor torques during operation, it is possible to extract information about friction (and wear) from available signals. Friction is however dependent on other factors than wear, such as **temperature** and **load**. The effects of temperature are specially difficult since temperature is not measured in typical robot applications. These effects should nevertheless be considered when evaluating different fault detection algorithms.

To simplify the presentation and due to confidentiality issues, the FDAs considered in the study are treated as black-boxes, processing data to generate a test quantity y, recall Fig. 1. The focus is placed on the FDAs *evaluation and comparison*. The FDAs considered in this study share the following characteristics, which are relevant for the presentation of the paper:

- C-1 Process data batches collected from a test-cycle;
- C-2 Outputs a scalar quantity for each data batch;
- C-3 Require nominal (wear-free) data.
- C-4 Process data for a single axis and should indicate wear changes only for that axis;
- C-5 The behavior of a test quantity depends on a combination of FDA, robot, axis and test-cycle;

Data is collected at Level 2, i.e. based on simulation experiments, using an ABB internal simulation tool and the analyses are performed at Level 3 with the use of surrogate models. A simplified version of the friction model presented in Bittencourt et al. [2011] is included in the

simulation model. The model used to describe friction in the robot joints is given by the static nonlinear function

$$\tau_f(\dot{\phi}, T, f) = \eta_0 + (\eta_1 + \eta_2 T)e^{-\left|\frac{\dot{\phi}}{\eta_3+\eta_4 T}\right|} + (\eta_5 + \eta_6 T)\dot{\phi} + \eta_7 e^{-\left|\frac{\dot{\phi}}{\eta_8 f}\right|} + \eta_9 \dot{\phi} f \quad (17)$$

and relates to the effects of angular speed, $\dot{\phi}$, temperature (as measured in the joint lubricant), T, and wear fault, f, to friction, τ_f. The remaining quantities $\boldsymbol{\eta}$ in (17) are model parameters, see Bittencourt and Gunnarsson [2012], Bittencourt et al. [2011] for more on friction models and their identification. The complete study includes:

S-1 Three FDAs (A,B,C) for wear diagnosis;
S-2 A medium robot with max. payload of 10-25kg,
S-3 A large robot with max. payload of 100-250kg;
S-4 Wear is studied in the three first axes of these robots.
S-5 A total of six different test-cycles.

The study is aimed at answering questions Q-1 to Q-3 for the robotics application. The next sections define the experiments performed and presents the results.

6.1 Design of Experiments

Input factors The following factors are considered relevant and are included in the study.

Wear. According to S-4, wear is introduced in three of the axes. Recalling that the FDAs process data for a single axis (C-4), the wear introduced in that axis will correspond to the fault factor f. When wear is present in the other two axes, they may cause variations in y due to coupling effects. Since these variations may complicate fault isolation, they are considered as nuisance factors, w_1 and w_2. The wear f in (17) is a dimensionless quantity with values between 0 (no wear) and 100 (a total failure due to wear), see Bittencourt et al. [2011]. In this study, it is considered that values in the range $[0, 50]$ are of interest. This is because the detection of a partial failure is more interesting for condition-based maintenance since it gives more time to perform maintenance before an eventual stop.

Temperature. The friction model used given in (17) includes temperature dependencies which will affect the data used for the FDAs. The temperature factor is assigned as w_3. The temperature range considered is $[30, 70]$°C and is based on a typical temperature behavior for a robot operating in a room with controlled environment temperature. The range copes with variations due to self-heating caused by losses in the joint and changes in the environment temperature.

Point-to-point delay. In point-to-point movements, the robot is required to fulfil a set of criteria in order to guarantee that a certain position was reached before issuing a command to move to the next position. During real-time path execution, the time required for the verification of these criteria may differ, causing variations to the trajectory. This varying "delay" is considered to have an effect on the test quantities and is thus included as a factor, w_4. The range of values for w_4 is $[25, 75]$ms and is based on values found for the robots studied.

Payload mass error. The control system used in the robot relies on the defined payload mass. The closed-loop system (and data) will thus be affected in case there is an error in the defined mass. The payload mass error is assigned as w_5 and has values in $[-10, 10]$% relative to the correct mass.

Table 1. Definition of the factor levels used.

	Coded Levels					
	-2	-1	0	1	2	
Factor	**Natural Levels**					**Unit**
f, w_1, w_2, wear	0	12.5	25	37.5	50	-
w_3, temperature	30	40	50	60	70	°C
w_4, point-to-point delay	25	37.5	50	62.5	75	ms
w_5, payload mass error	-10	-5	0	5	10	%

Table 2. Some entries of the design matrix.

Row	f	w_1	w_2	w_3	w_4	w_5		...					
1	-1	-1	-1	-1	-1	-1	41	0	0	0	0	-2	0
2	-1	-1	-1	-1	1	1	42	0	0	0	0	2	0
3	-1	-1	-1	1	-1	1	43	0	0	0	0	0	-2
4	-1	-1	-1	1	1	-1	44	0	0	0	0	0	2
5	-1	-1	1	-1	-1	1	45	0	0	0	0	0	0

For the study, five levels are considered for each factor, i.e. $m = 5$. The levels are distributed linearly within the suitable range for the factors. The factor levels used can be seen in Table 1. According to C-3, the test quantities require nominal (wear-free) data which are generated according to the following coded levels,

$$\mathbf{z}^0 = [\,-2\ -2\ -2\ 0\ 0\ 0\,]^T, \quad (18)$$

i.e. no wear is present in any of the axes, with temperature at 50°C, 50 ms of point-to-point delay and no error in payload mass.

Regression Models Two model structures are considered, a full second-order model as in (3) and a simplified second order model of the form

$$\varphi(\mathbf{z})^T = [\,1\ f\ \mathbf{w}^T\ \text{svec}(\mathbf{w}\mathbf{w}^T)^T\,] \quad (19a)$$

$$\boldsymbol{\theta}^T = [\,b\ \theta_0\ \{\theta_i\}\ \{\theta_{ij}\}\ \{\theta_{ii}\}\,] \quad (19b)$$

where $i, j \in [1, \cdots, n]$ with $i > j$. Notice that there are no cross-terms for the fault f in model structure (19), only for the disturbances $\mathbf{w}$. An interpretation of the coefficients for this model is thus simpler compared to the full second-order model.

Design Matrix A central composite design based on a 2^{n-1} fractional factorial design with one center point and star points at [2,-2] is considered, requiring a total of $N = 45$ experiments. Parts of the values for the design matrix are seen in Table 2.

Design Parameters According to C-5, the test quantities produce comparable results only when the same FDA, robot, axis and test-cycle is used. Therefore, for different combinations of these *design parameters*, a different regression model should be used. This gives a total of $3 \times 2 \times 3 \times 6 = 108$ regression models corresponding to the number of FDAs, robots, axes and test-cycles considered. Notice though that the same design matrices can be used to identify all regression models. And further that the *same simulated data* for a robot can be used to identify the models for all FDAs and for all axes. Each regression model requires $N = 45$ experiments, a total of $108/(3{\times}3){\times}45 = 540$ simulations are therefore needed to identify all regression

Table 3. Factor levels used for validation.

Factor	Natural Levels					Unit
f, w_1, w_2, wear	5	15	25	35	45	-
w_3, temperature	35	42	49	56	63	°C
w_4, point-to-point delay	31	41	51	61	71	ms
w_5, payload mass error	-9	-6	-3	0	3	%

Table 4. Model fits for a robot and test-cycle.

		Model Fit [%]		
FDA	Model Eq.	Axis 1	Axis 2	Axis 3
A	(19)	83.2	72.5	82.1
A	(3)	87.9	83.7	88.1
B	(19)	64.6	65.5	65.7
B	(3)	87.8	91.0	91.9
C	(19)	89.8	84.2	85.6
C	(3)	95.0	85.3	89.9

models in the study. Each simulation experiment takes around ten seconds to be performed, requiring 90 min for all 540 simulations.

Identification and Validation The simulation experiments are performed and the regression models are identified using (5). The design matrix used for identification, given in Table 2, is also used for validation of the regression models but with different factor levels, given in Table 3. The model fits, computed as in (6), are shown in Table 4 for a certain robot and test-cycle. The fits are generally high for all FDAs, with higher values for the full second-order model, specially for FDA B.

6.2 Determining Relevant Factors

Sensitivity analysis are used here to address Q-1, i.e. to determine which input factors cause more variations to the output of an FDA. The regression coefficients are normalized as in (8) and are grouped for each FDA according to (9). Because model (19) is simpler to analyze than model (3), only the coefficients for this model are shown here. Each of the group matrices, as in (9), have dimensions (36×21), corresponding to a combination of the design parameters left and number of coefficients in the regression model.

In Fig. 2, the 21 normalized coefficients are displayed in box plots for each test quantity. The statistics for the box plots are computed over each column of the group matrices. Recall that, because of the normalization used, coefficients with values larger than 1 indicate that the corresponding regression term has a larger effect to the output compared to the direct effect of f, i.e. wear. An inspection of the graphs allows to conclude that factor w_3, temperature, considerably affects all FDAs. FDA C has the lowest value for the median of coefficient $\bar{\theta}_3$, related to the direct influence of temperature. For FDA A, values greater than 1 are found for $\bar{\theta}_5$, the direct influence of payload mass error, while FDAs B and C show a less significant response for this factor. The factors related to wear in other joints, w_1 and w_2, and point-to-point delay w_4 show insignificant responses.

6.3 Comparing Fault Detection Algorithms

As discussed in Sec. 4, the SNRs can be seen as a measure of average performance of a test quantity. For the computation of the SNRs, the parameters defining the hypotheses in (10) are set as

$$f^0 = -0.5, \quad \Delta = 1, \quad p(\mathbf{w}) = \mathcal{N}(0, I\sigma^2), \quad \sigma = 0.25, \quad (20)$$

i.e. the main wear factor, f, is changed by a fourth of its allowed range and the nuisance factors $\mathbf{w}$ are considered as Gaussian random variables independently distributed with a common standard deviation which is 1/16 of their range. Model structure (3) is considered in the study since it presented larger fits in general (recall Table 4). For each regression model, the SNRs are computed based on $1\,10^5$ MC runs. The total $1.08\,10^7$ MC runs needed for all regression models took approximately 12 seconds in a standard desktop PC. To perform the same analysis using Level 2 studies, i.e. with the simulator, would have taken nearly three and a half years.

The SNRs grouped according to FDA are displayed in box plots in Fig. 3. The SNRs can be used to rank the different FDAs. If the median over each group is used as a criterion, this example reveals that FDA C gives better performance, followed by FDAs B and A.

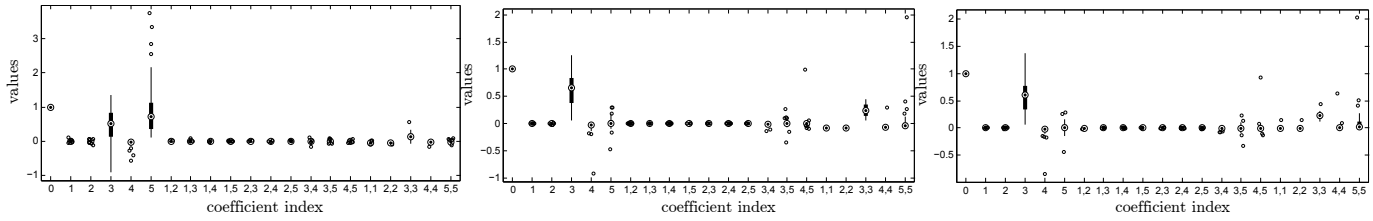
6.4 Determining the Effective Scope

The use of scope matrices is illustrated here to determine how the factors w_3 and w_5, i.e. temperature and payload mass error, delimit the scope of the test quantities. Criterion (12) is considered for the study with $P'_f = 0.01$ and $P'_d = 0.99$. The hypotheses in (15) are defined with $f^0 = -2$ and $w'_{j \neq i}$ are set to the nominal values given in (18). The criterion is evaluated for values of $\Delta \in [0, 4]$ and $\sigma \in [0.01, 1]$ based on a linear grid of size 30×30. The hypotheses densities are estimated using a kernel density estimator based on $1\,10^5$ MC runs. The total MC runs needed for the study is of $30{\times}30{\times}1.08\,10^7 = 9.72\,10^9$ which took approximately 3h15min using the regression models. To evaluate these analyses at Level 2, with the simulator, would have taken more than three millennia.

Group scope matrices are formed for each FDA. An entry in the resulting matrix can take values between zero and 36. The resulting matrices for w_3 and w_5 are shown in Fig. 4 with a colormap associated to the entry value in the scope matrix. From an inspection of the figures, it is possible to determine the minimal size of Δ for which an FDA performs satisfactorily given a fixed disturbance variation σ, and vice-versa. From analyses of Fig. 4(b), it is possible to note that FDA C is the least affected by payload disturbances. As seen in Fig. 4(a), all test quantities are considerably affected by temperature, but FDAs A and C allow for more variations of temperature compared to FDA B.

7. CONCLUSIONS

This paper proposed a framework for evaluation and comparison of fault detection algorithms (FDAs) based on simulations. An extensive investigation of the different FDAs is made possible with the use of surrogate models which considerably reduce the time needed for the evaluation of the analyses. As illustrated in the application example, this was in fact the only viable alternative. The approaches suggested may be used to reveal which inputs affect an



(a) FDA A. (b) FDA B. (c) FDA C.

Fig. 2. Normalized regression coefficients for the model (19) grouped according to FDA. In the box plots, the dotted circle indicates the median, the extremities of the bar relate to the 25th and 75th percentiles and the isolated circles are outliers. Notice the different scales.

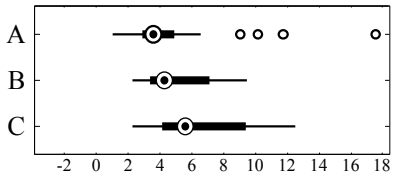
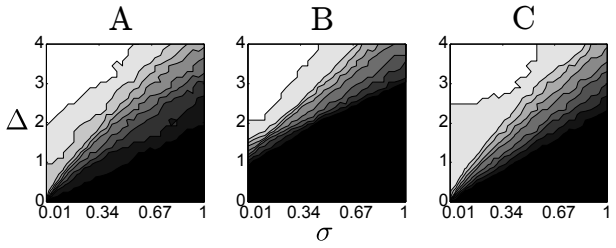


Fig. 3. SNRs grouped for the different FDAs. The box plots are for groups over all design parameters where the dotted circle indicates the group median, the extremities of the bar relate to the 25th and 75th percentiles and circles are outliers.

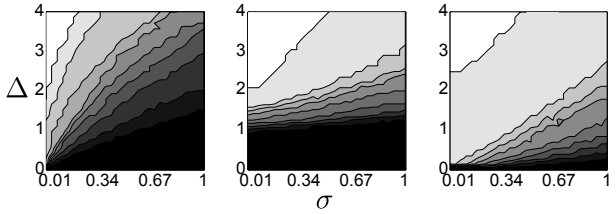
FDA the most, which FDA performs best in average and the effective scope of an FDA.

It should be noted that conclusions drawn based on simulations or surrogate models should always be carried out carefully since they are a limited representation of reality. Results achieved in this manner give good insights about the problem and support decisions but, ultimately, the test quantities should be evaluated based on real experiments. In the robotics application, accelerated wear tests can be used with this purpose, but with much higher costs and time required for a statistically significant study.

The framework is rather general and can be extended to study various aspects of fault detection algorithms. For example, *tuning parameters and sampling frequency* could be included in $\mathbf{z}$ to study how they affect the test quantities and support their choices.



(a) Random temperature disturbances.



(b) Random payload mass error disturbances.

Fig. 4. Visualization of the scope matrices grouped according to FDA. The colormap relates to how often the performance test was successful, varying from 0 (black) to 36 (white). The clearer the plot, the more often a FDA performed satisfactorily.

REFERENCES

A. Atkinson, A. Donev, and R. Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, Cary, USA, 2007.

A. C. Bittencourt and S. Gunnarsson. Static friction in a robot joint—modeling and identification of load and temperature effects. *Journal of Dynamic Systems, Measurement, and Control*, 134(5), 2012.

A. C. Bittencourt, P. Axelsson, Y. Jung, and T. Brogårdh. Modeling and identification of wear in a robot joint under temperature disturbances. In *Proc. of the 18th IFAC World Congress*, Milan, Italy, Aug 2011.

A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series)*. Oxford University Press, USA, Nov. 1997.

G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. John Wiley and Sons, 1978.

V. V. Fedorov. *Theory of optimal experiments*. Academic Press, 1972.

P. Frank and X. Ding. Survey of robust residual generation and evaluation methods in observer-based fault detection systems. *Journal of Process Control*, 7(6):403 – 424, 1997.

F. Gustafsson. *Adaptive Filtering and Change Detection*. Wiley, Oct. 2000.

R. Isermann. On the applicability of model-based fault detection for technical processes. *Control Engineering Practice*, 2(3):439 – 450, 1994. ISSN 0967-0661.

R. Isermann. *Fault-Diagnosis Systems - An Introduction from Fault Detection to Fault Tolerance*. Springer, 2006.

J. P. C. Kleijnen, S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. A User's Guide to the Brave New World of Designing Simulation Experiments. *INFORMS Journal on Computing*, 17, 2005.

X. Li and K. Zhou. A time domain approach to robust fault detection of linear time-varying systems. *Automatica*, 45(1):94 – 102, 2009.

N. Liu and K. Zhou. Optimal robust fault detection for linear discrete time systems. *J. Control Sci. Eng.*, 2008:7:1–7:16, January 2008.

L. Ljung. *System Identification: Theory for the User (2nd Edition)*. Prentice Hall PTR, December 1998.

J. E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769, 2004.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Carboni, M. S. D. Gatelli, and S. Tarantola. *Global Sensitivity Analysis. The Primer*. John Wiley and Sons, 2008.

S. Sanchez. Work smarter, not harder: guidelines for designing simulation experiments. In *Proc. of the 2006 Winter Simulation Conference*, Monterey, USA, dec. 2006.

H. L. Van Trees and L. B. Kristine. *Detection, Estimation and Modulation Theory, Part I*. Wiley, New York, USA, 2nd edition, 2013.

X. Wei and M. Verhaegen. Robust fault detection observer design for linear uncertain systems. *International Journal of Control*, 84 (1):197–215, 2011.