Automatic tuning via Kriging-based optimization of methods for fault detection and isolation

Julien Marzat, Eric Walter, Hélène Piet-Lahanier, Frédéric Damongeot

Automatic tuning via Kriging-based optimization of methods for fault detection and isolation

Julien Marzat, Éric Walter, Hélène Piet-Lahanier, Frédéric Damongeot

***Abstract*— All the methods for Fault Detection and Isolation (FDI) involve internal parameters, often called *hyperparameters*, that have to be carefully tuned. Most often, tuning is ad hoc and this makes it difficult to ensure that any comparison between methods is unbiased. We propose to consider the evaluation of the performance of a method with respect to its hyperparameters as a *computer experiment*, and to achieve tuning via global optimization based on Kriging and Expected Improvement. This approach is applied to several residual-evaluation (or change-detection) algorithms on classical test-cases. Simulation results show the interest, practicability and performance of this methodology, which should facilitate the automatic tuning of the hyperparameters of a method and allow a fair comparison of a collection of methods on a given set of test-cases. The computational cost turns out to be much lower than the one obtained with other general-purpose optimization methods such as genetic algorithms.**

***Index Terms*— hyperparameter, method adjustment, parameter tuning, residual evaluation, change detection, fault detection and isolation, efficient global optimization, Kriging**

I. Introduction

A fault detection and isolation (FDI) procedure is usually made up of a residual generator, and a method for residual analysis that processes these residuals [1]. This is used to decide whether a fault is present and then which fault.

Each of the many change-detection methods has internal parameters that must be carefully tuned. These parameters, often called *hyperparameters*, have a strong impact on performance and robustness. The user may thus be at a loss for selecting the most efficient method. This can be achieved by first defining a suitable performance criterion and then finding a way of tuning the hyperparameters of each method in order to optimize this criterion, on a representative set of test-cases.

The main existing tools for the tuning of hyperparameters are *cross-validation* and its variants (k-fold cross-validation, leave-one-out cross-validation, generalized cross-validation [2]). Cross-validation is used to estimate the performance for a given value of the hyperparameter vector and can then be complemented by an optimization procedure to find the best tuning of these hyperparameters. In [3] and [4], such approaches based on a discretization of hyperparameter spaces have been presented. Another method using Bayesian networks for tuning parameters has been proposed in [5], where prior knowledge consists of the previous simulation runs. All these approaches prove to be extremely computer-intensive and thus not applicable when the simulation budget is severely limited.

This paper describes an optimization procedure that is dedicated to this type of problem, and its application to the automatic tuning of methods for FDI. Following the *computer experiment* framework [6], we propose to use a global optimization algorithm relying on *Kriging* and the notion of *Expected Improvement* to explore real-valued hyperparameter spaces effectively at a limited computational cost.

Section II formally presents the problem and explains the basics of the tuning methodology. Section III describes illustrative test-cases, examples of candidate methods to be tuned and compared, along with classical performance indices in FDI to be used as optimization criteria. Results are reported in Section IV, and conclusions and perspectives in Section V.

II. Hyperparameter-tuning methodology

A. Problem formulation

Assume several FDI methods compete for the same application. The j-th method depends on a vector $\mathbf{x}^j \in \mathbb{X}^j \subset \mathbb{R}^{d_j}$ of hyperparameters, where $\mathbb{X}^j$ is the feasible hyperparameter space and $d_j = \dim \mathbf{x}^j$. All of these methods are to be compared using the same real-valued performance criterion y. This criterion could combine several performance indices, e.g., the trade-off between false-alarm and non-detection rates for change-detection procedures. Tuning the j-th method means looking for the value of $\mathbf{x}^j$ that minimizes $y\left(\mathbf{x}^j\right)$. A possible way to compare methods is then to rank them according to their best value for y.

The tuning of a given method is central to the selection of the best of them and will now be considered. For the sake of simplicity, the index j will be omitted in what follows. The cost function is thus a scalar function $y(\mathbf{x})$, where $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^d$. The only available information is the result of previous computer experiments that provides the value of $y(\mathbf{x})$ for given values of $\mathbf{x}$. The procedure is recursive and we shall assume that we have already computed n samples forming the vector $\mathbf{y}_n = [y_1, ..., y_n]^{\mathrm{T}}$ corresponding to $\mathcal{X}_n = [\mathbf{x}_1, ..., \mathbf{x}_n]$. Since the evaluation of $y(\mathbf{x})$ is expensive, we shall use a simpler prediction $\widehat{y}(\mathbf{x})$ of $y(\mathbf{x})$ based on these samples and obtained by *Kriging.*

B. Basics of Kriging

Kriging has been given this name by the French geostatistician G. Matheron, to recognize the seminal influence of

J. Marzat, H. Piet-Lahanier and F. Damongeot are with ONERA-DPRS, Palaiseau, France, firstname.lastname@onera.fr

J. Marzat and É. Walter are with the Laboratoire des Signaux et Systèmes (L2S), CNRS-SUPELEC-Univ-Paris-Sud, France, firstname.lastname@lss.supelec.fr

the work of D.G. Krige on the gold deposit of the Rand, in South Africa [7]. In Kriging, the function $y(\cdot)$ is modeled as a Gaussian process $Y(\cdot)$ with mean function $m(\cdot)$ and covariance function $k(\cdot,\cdot)$ [8]. More specifically, $Y(\cdot)$ is written as

$$Y(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\,\mathbf{b} + Z(\mathbf{x})$$

where $\mathbf{f}(\mathbf{x})$ is some known regression function vector (usually chosen constant or polynomial in $\mathbf{x}$), $\mathbf{b}$ is a vector of unknown regression coefficients to be estimated, and $Z(\cdot)$ is a zero-mean Gaussian process with known (or parametrized) covariance function $k(\cdot,\cdot)$. Kriging is then the search for the *best linear unbiased predictor* (BLUP) of $Y(\cdot)$ [9].

The actual covariance $k(\cdot,\cdot)$ is usually unknown. It is expressed as

$$k\left(Z(\mathbf{x}_i), Z(\mathbf{x}_j)\right) = \sigma_Z^2 R(\mathbf{x}_i, \mathbf{x}_j)$$

where σ_Z^2 is the process variance and $R(\cdot,\cdot)$ is a parametric correlation function. Both σ_Z^2 and the parameters of $R(\cdot,\cdot)$ must be chosen or estimated from the available data. Under a stationarity assumption, $R(\mathbf{x}_i, \mathbf{x}_j)$ depends only on the displacement vector $\mathbf{x}_i - \mathbf{x}_j$, denoted by $\mathbf{h}$ in what follows. A frequent choice of correlation function, also adopted in the present paper, is the *power exponential correlation function*

$$R(\mathbf{h}) = \exp\left(-\sum_{k=1}^{d}\left|\frac{h_k}{\theta_k}\right|^{p_k}\right)$$

where $0 < p_k \leq 2$, and h_k is the k-th component of $\mathbf{h}$. Note that with this choice, $R(\mathbf{h})$ tends to 1 when $\mathbf{h}$ tends to $\mathbf{0}$. The θ_k may be estimated from the data by maximum likelihood, to get what is known as *empirical Kriging* (this setting has been used for the application reported in Section IV). A wide range of other choices for the correlation function is available [6].

Define $\mathbf{R}$ as the $n \times n$ matrix such that

$$\mathbf{R}(i,j) = R(\mathbf{x}_i, \mathbf{x}_j)$$

$\mathbf{r}(\mathbf{x})$ as the n vector

$$\mathbf{r}(\mathbf{x}) = \left[R(\mathbf{x}, \mathbf{x}_1), ..., R(\mathbf{x}, \mathbf{x}_n)\right]^{\mathrm{T}}$$

and $\mathbf{F}$ as the $n \times \dim \mathbf{b}$ matrix

$$\mathbf{F} = \left[\mathbf{f}(\mathbf{x}_1), ..., \mathbf{f}(\mathbf{x}_n)\right]^{\mathrm{T}}$$

In this presentation, we assume, for the sake of simplicity, that the parameters of the covariance matrix are known, but remember that in our application they are estimated by maximum likelihood. The maximum-likelihood estimate $\widehat{\mathbf{b}}$ of the regression coefficients $\mathbf{b}$ from the available data $\{\mathcal{X}_n, \mathbf{y}_n\}$ is

$$\widehat{\mathbf{b}} = \left(\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}_n$$

The predictor of the mean of the Gaussian process, at $\mathbf{x} \in \mathbb{X}$, is then given by

$$\widehat{Y}(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\,\widehat{\mathbf{b}} + \mathbf{r}(\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y}_n - \mathbf{F}\widehat{\mathbf{b}}\right)$$

This predictor is linear in $\mathbf{y}_n$ and interpolates the training data, as $\widehat{Y}(\mathbf{x}_i) = y_i$. Another interesting property of Kriging, which is crucial regarding global search, is the possibility to compute the *variance of the prediction error* [10] at $\mathbf{x} \in \mathbb{X}$ by

$$\widehat{\sigma}^2(\mathbf{x}) = \sigma_Z^2\left(1 - \mathbf{r}(\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}(\mathbf{x})\right)$$

C. Maximizing Expected Improvement

The idea is to use the Kriging predictor $\widehat{Y}$ to find the $(n+1)$-st point at which a simulation of the complete FDI process will be run. This point is chosen according to a criterion $J(\cdot)$ that measures the interest of an additional evaluation at $\mathbf{x}$, given the past results $\mathbf{y}_n$ obtained at $\mathcal{X}_n$ and the Kriging prediction of the mean $\widehat{Y}(\mathbf{x})$ and variance $\widehat{\sigma}^2(\mathbf{x})$,

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in\mathbb{X}} J\left(\mathbf{x}, \mathcal{X}_n, \mathbf{y}_n, \widehat{Y}(\mathbf{x}), \widehat{\sigma}^2(\mathbf{x})\right)$$

A common choice for $J(\cdot)$ is *Expected Improvement* [11]. The best available estimate of the minimum of y after the first n evaluations is $y_{\min}^n = \min_{i=1\ldots n}\{y_i = y(\mathbf{x}_i)\}$. With

$$u = \left(y_{\min}^n - \widehat{Y}(\mathbf{x})\right)/\widehat{\sigma}(\mathbf{x})$$

the Expected Improvement is expressed in closed-form as

$$\mathrm{EI}(\mathbf{x}) = \widehat{\sigma}(\mathbf{x})\left[u\Phi(u) + \phi(u)\right]$$

where Φ is the cumulative distribution function and ϕ the probability density function of the normalized Gaussian distribution $\mathcal{N}(0,1)$. Maximizing Expected Improvement achieves a trade-off between local search (numerator of u) and the exploration of unknown areas (where $\widehat{\sigma}$ is high) and is therefore well suited for global optimization.

D. EGO algorithm

The global optimization procedure that has been used for this study, based on the aforementioned elements, is called EGO, for *efficient global optimization* [12]. A preliminary sampling is required to obtain the n points of the initial design $\mathcal{X}_n$. Latin Hypercube Sampling (LHS) has been chosen to explore $\mathbb{X}$ evenly [13]. The description of EGO is given in Algorithm 1. The algorithm stops either when the maximal number of iterations $n_{\max}$ (which depends on the simulation budget) is reached or when the Expected Improvement becomes lower than some threshold ϵ. Our implementation is based on Sasena's toolbox SuperEGO [14] and uses the DIRECT optimization algorithm [15] to achieve Step 5 of Algorithm 1.

III. Illustrative application to the choice of a residual-evaluation strategy

This section presents the residual-analysis methods that will be tuned and compared, performance indices as goals for the optimization procedure and two classical test-cases. It should be noted that the methodology advocated in this paper can be applied to a much broader class of problems, and that the selection considered here is just for the purpose of illustration. Indeed, EGO is particularly well suited to problems where the evaluation of y is computationally expensive, as would be the case, for instance, when using cross-validation.

Algorithm 1: EGO

Choose $\mathcal{X}_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ by LHS in $\mathbb{X}$
Compute $\mathbf{y}_n = \{y(\mathbf{x}_1), ..., y(\mathbf{x}_n)\}$
while $\max_{\mathbf{x}\in\mathbb{X}} \{\mathrm{EI}(\mathbf{x})\} > \varepsilon$ ***and*** $n < n_{\max}$ **do**
    Fit the Kriging model on the known data points $\{\mathcal{X}_n, \mathbf{y}_n\}$ as described in Section II-B
    Find $y^n_{\min} = \min_{i=1...n} \{y(\mathbf{x}_i)\}$
    Find the next point of interest $\mathbf{x}_{n+1}$ by maximizing Expected Improvement as described in Section II-C
    Compute $y(\mathbf{x}_{n+1})$, append it to $\mathbf{y}_n$ and append $\mathbf{x}_{n+1}$ to $\mathcal{X}_n$
    $n \leftarrow n + 1$
end

A. Strategies to be evaluated

A scalar residual $r(t)$ is a signal that should remains negligible as long as there is no fault to which it is sensitive, and that becomes sufficiently large to be noticeable when a fault occurs. We consider residual-evaluation methods that provides a scalar binary decision function, which should return *false* if the residual is close enough to its initial mean (usually zero) and *true* if a jump or a drift occurs in the signal. The problem to be solved here is to detect a statistical change in the mean from its initial value zero to an unknown but different value.

Six candidate methods are to be tuned and compared by the proposed methodology. The operating principle of each of them is briefly recalled to highlight the hyperparameters involved, and references are given for further details. As the nominal mean μ_0 and variance σ_0^2 of the signal are usually required, we estimate them on the first data for all methods and do not include them in the hyperparameters to be tuned.

1) The "three sigma" rule: This method proposes to choose bilateral fixed thresholds equal to $\mu_0 \pm \nu\sigma_0$, where $\nu \geq 3$ usually [16], relying on the fact that 99.7% of the points of a Gaussian distribution lie within three standard deviations. The decision takes the value *true* when the value of the residual falls outside the thresholds, else the decision is *false.*

2) Student's t-test: This test checks whether the signal follows a Gaussian distribution $\mathcal{N}(\mu_0, \sigma_0)$, which leads to an automatic thresholding given by Student's table considering that the required confidence level is fixed here at 5% [17]. The test is applied to a sliding window of width N.

3) Generalized Likelihood Ratio (GLR) test: This test is based on the likelihood ratio $\Lambda(r)$ of the probability that the mean of r is $\mu_1 \neq \mu_0$ to the probability that it is μ_0, still assuming that the signal is Gaussian [18], [19]. The generalized version uses the maximum-likelihood estimate $\widehat{\mu}_1$ of μ_1 to allow the detection of a change of unknown magnitude. The practical implementation using a sliding window of width N and the log-likelihood ratio is given by

$$\begin{cases} \sum_{t=1}^{N} r(t) > \frac{\sigma_0^2}{\widehat{\mu}_1 - \mu_0} \ln(\lambda) + \frac{N(\mu_0 - \widehat{\mu}_1)}{2} \Longrightarrow \text{decide } true \\ \text{else} \Longrightarrow \text{decide } false \end{cases}$$

where the threshold λ is one of the hyperparameters.

4) Sequential Probability Ratio Test (SPRT): The SPRT is very similar to the GLR, as it also uses the likelihood ratio on a sliding window of width N. However, the minimum change detection size μ_1 has to be specified, and the threshold λ is determined by the desired false-alarm and non-detection probabilities, respectively α and β [19]. The following decisions are taken at each step:

$$\begin{cases} \Lambda < \frac{\beta}{1-\alpha} & \Longrightarrow \text{decide } false \\ \Lambda > \frac{1-\beta}{\alpha} & \Longrightarrow \text{decide } true \\ \text{else} & \text{take no decision} \end{cases}$$

5) CUSUM test : No statistical hypothesis is needed here. This two-sided test is expressed as follows [19][20]

$$\begin{cases} S_1(t) = \max(S_1(t-1) + r(t) - \mu_0 - \delta/2, 0) \\ S_2(t) = \max(S_2(t-1) - r(t) + \mu_0 - \delta/2, 0) \end{cases}$$

where δ is the minimal size of the fault to be detected. The decision rule is then

$$\begin{cases} (S_1 > \lambda) \text{ or } (S_2 > \lambda) \Longrightarrow \text{decide } true \\ \text{else} \Longrightarrow \text{decide } false \end{cases}$$

where the threshold λ is one of the hyperparameters.

6) Randomised SubSampling (RSS): This very recent method, proposed in [21], uses M subsamplings of the signal on a sliding window of width N. The sum of the errors with respect to the expected mean μ_0 is computed on each subsample. The decision is *false* if at least q of the M sums are greater than zero and at least q of the M sums are smaller than zero, else the decision is *true*. An interesting property of the test is that the expected probability of false alarm is $\alpha_{\exp} = 2q/M$.

Table I summarizes the hyperparameters involved in the methods considered.

TABLE I: Hyperparameters of the candidate methods

3–Sigma	Student	GLR	SPRT	CUSUM	RSS
ν	N	N, λ	N, μ_1, α, β	δ, λ	N, q, M

B. Performance indices

We propose to use some of the quantitative indices defined within the DAMADICS benchmark [22]. Figure 1 shows time zones in the evolution of the Boolean decision function that are the basis of the definition of the performance indices. The value of the function before t_{on} and after t_{hor} is not to be taken into account, while t_{from} is the instant at which the fault occurs. The indices that will be used for performance evaluation are

- the *detection delay* t_{dt}, which is the time elapsed between the fault occurrence time t_{from} and the last instant of time at which the decision signal switched from *false* to *true*;

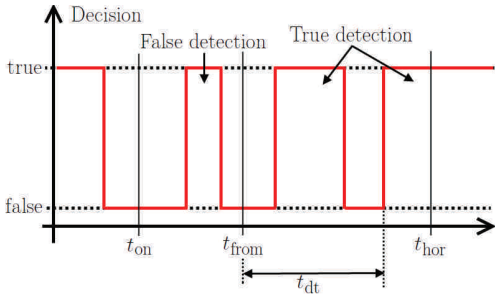


Fig. 1: Time zone parameters for the definition of performance indices

- the *false-detection rate* $r_{\text{fd}} = \left(\sum_i t^i_{\text{fd}}\right) / \left(t_{\text{from}} - t_{\text{on}}\right)$, where t^i_{fd} is the i-th period of time between t_{on} and t_{from} where the decision is *true*;
- the *non-detection rate* $r_{\text{nd}} = 1 - r_{\text{td}}$, where $r_{\text{td}} = \left(\sum_i t^i_{\text{td}}\right) / \left(t_{\text{hor}} - t_{\text{from}}\right)$ is the *true-detection rate* with t^i_{td} the i-th period of time between t_{from} and t_{hor} where the decision is *true*.

C. Test-cases

The classical test-cases [19], [20], [21] that will be used correspond to a Gaussian signal with unit variance and a signal uniformly distributed on $[-2; 2]$. Both signals consist of 1000 points with a jump in the mean from 0 to 1 at $t_{\text{from}} = 500$, with $t_{\text{on}} = 0$ and $t_{\text{hor}} = 1000$ (see Figure 2). They have been generated with a seed equal to 7361731 in Matlab.

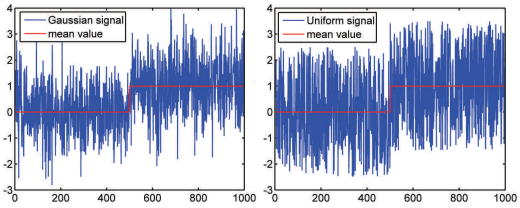


Fig. 2: Gaussian (left) and Uniform (right) test-cases

IV. Results

A. Setting

The initial sampling consists of an LHS of $10d$ points ($d = \dim \mathbf{x}$), as suggested in [12]. The nominal mean and variance of the signals are estimated on the first 100 data points. Stopping parameters are $n_{\max} = 100$ and $\varepsilon = 10^{-4}$. This means that 100 simulations are to be run at most and prove to be most of the time not even necessary. This is a clear advantage of Kriging-based optimization, as evolutionary algorithms would typically require many thousands evaluations.

The cost function of the global optimization problems considered by EGO is scalar. The simplest way to achieve multiobjective optimization with the performance indices defined in Section III-B is to minimize some weighted global cost function $c = w_{\text{fd}} r_{\text{fd}} + w_{\text{nd}} r_{\text{nd}} + w_{\text{dt}} t_{\text{dt}}$ where the $w_{(.)}$s are positive weights to be chosen. As the two indices r_{fd} and r_{nd} take values in $[0; 1]$, the weights w_{fd} and w_{nd} can be taken equal to 1, for an unprejudiced trade-off. The detection delay could also be included in the criterion, but should be normalized to match the range of the two other indices. Two continuous cost functions have been used in this study, $c_1 = r_{\text{fd}} + r_{\text{nd}}$ and $c_2 = r_{\text{fd}} + r_{\text{nd}} + 0.01 \cdot t_{\text{dt}}$. The first one achieves the trade-off between false-detection and non-detection without taking explicitly delay into account, unlike the second one that also seeks for a reduced detection delay.

The feasible hyperparameter search spaces for all methods are indicated in Table II. Note that N, q and M are integers.

B. Results

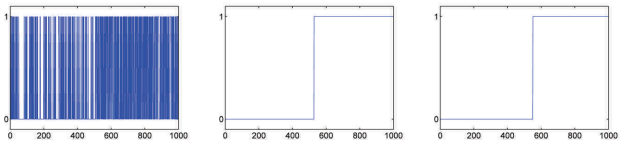
The tuning results obtained on the two test-cases with the cost functions c_1 and c_2 for the candidate methods are presented in Tables III, IV, V and VI. The optimal values of the cost and the corresponding ranking of the methods are given, along with the values taken by the three performance indices from Section III-B and the corresponding hyperparameter tuning. Figures 3 and 4 show the decision functions corresponding to the best setting for each method on both test-cases. Explorations of the hyperparameter spaces (those with no more than two hyperparameters) by the global-optimization algorithm EGO are displayed on Figure 5. An acceptable tuning has been successfully found for each method, within $n_{\max}$ runs of the simulation. Although the examples treated here contain no more than four hyperparameters, nothing in the method forbids considering higher-dimensional problems.

Even if these two test-cases are not sufficient to assess the absolute ability of these methods, some trends can be spotted. It appears that the 3-sigma method is not well suited to detect a change of the same order of magnitude as the standard deviation of the signal. Student's test and the GLR test perform better if the Gaussian hypothesis stands true. The best results have been obtained with the SPRT test, the RSS approach and especially the CUSUM test. A possible explanation is that the latter two tests are not based on statistical hypothesis and only require the noise to be symmetrically distributed around the mean.

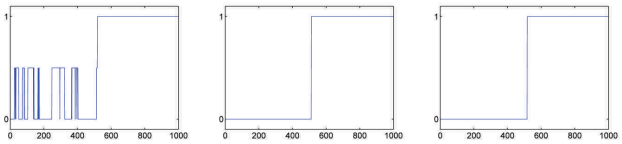
The two criteria often (but not always) yield similar results. This is due to the complementary goal shared by the minimization of t_{dt} and r_{nd}. To check the sensitivity of the results to the choice of the initial LHS, we ran the EGO algorithm several times with randomly chosen initial samples. The results proved to be quite robust to initialization and none of them falsified the conclusions presented here (e.g., 250 runs for Student tuning with c_1 gave a mean of 0.0561 with standard deviation of $4.5 \cdot 10^{-7}$ for the best cost).

V. Conclusions and Perspectives

We have presented a methodology based on computer experiment and Expected Improvement techniques for tuning the hyperparameters of all the approaches that we wish to compare. The methodology is applicable to any parameter tuning problem, assuming that a computer simulation of the

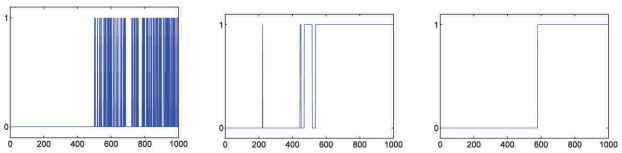


(a) 3-sigma (b) Student (c) GLR

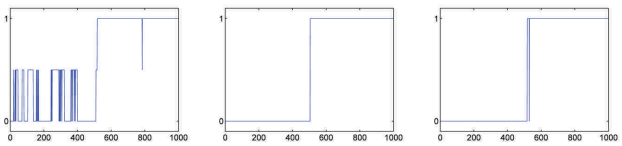


(d) SPRT (no decision: 0.5) (e) CUSUM (f) RSS

Fig. 3: Decision functions on the Gaussian test-case



(a) 3-sigma (b) Student (c) GLR



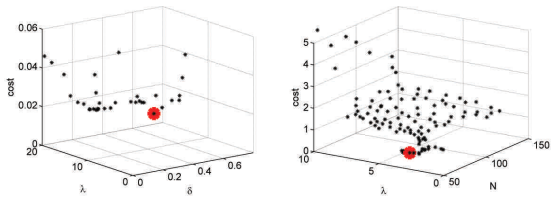
(d) SPRT (no decision: 0.5) (e) CUSUM (f) RSS

Fig. 4: Decision functions on the Uniform test-case

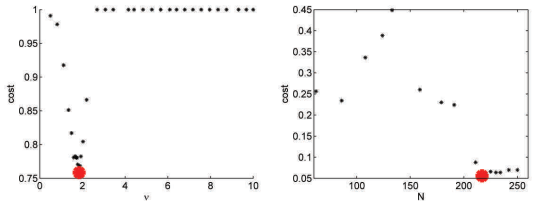
problem is available and that performance indices are computable. Kriging acts as a surrogate and simple-to-compute approximation of the complicated simulation leading to the evaluation of the performance indices. A global optimization procedure using the Kriging predictor then looks for the best real-valued hyperparameters.

The practicability of the methodology has been successfully illustrated through the selection of a residual-analysis strategy among various change-detection methods. Future work will address the evaluation of whole diagnosis strategies, comprising a residual generator coupled with an analysis algorithm on representative case-studies. These methods will necessarily imply more hyperparameters and the practical applicability of the method to larger dimensions will therefore be addressed. As a more general FDI case-study will involve model and measurement uncertainty, there is also the need to take into account *environmental variables* [6] (time of occurrence of faults, noise level, model uncertainty level...). Other multiobjective optimization techniques may also be investigated.

This paper employed the most classical method for Kriging-based global optimization, namely EGO. Alternative approaches, such as IAGO [23] could also be considered.



(a) CUSUM (b) GLR



(c) 3-sigma (d) Student

Fig. 5: Exploration of some hyperparameter spaces by EGO ; best tuning is in red

REFERENCES

[1] R. Isermann, "Supervision, fault-detection and fault-diagnosis methods: An introduction," *Control Engineering Practice*, vol. 5, no. 5, pp. 639–652, 1997.

[2] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[3] R. Kohavi and G. H. John, "Automatic parameter selection by minimizing estimated error," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 304–312.

[4] F. Hutter, H. H. Hoos, and T. Stutzle, "Automatic algorithm configuration based on local search," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 22, no. 2, 2007, pp. 1152–1160.

[5] R. Pavón, F. Díaz, and V. Luzón, "A model for parameter setting based on Bayesian networks," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 1, pp. 14–25, 2008.

[6] T. J. Santner, B. J. Williams, and W. Notz, *The design and analysis of computer experiments*. Springer-Verlag, Berlin-Heidelberg, 2003.

[7] G. Matheron, "Principles of geostatistics," *Economic Geology*, vol. 58, no. 8, p. 1246, 1963.

[8] J. Lefebvre, H. Roussel, E. Walter, D. Lecointe, and W. Tabbara, "Prediction from wrong models: the Kriging approach," *IEEE Antennas and Propagation Magazine*, vol. 38, no. 4, pp. 35–45, 1996.

[9] J. P. C. Kleijnen, "Kriging metamodeling in simulation: A review," *European Journal of Operational Research*, vol. 192, no. 3, pp. 707–716, 2009.

[10] M. Schonlau, *Computer Experiments and Global Optimization*. PhD thesis, University of Waterloo, Canada, 1997.

[11] D. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of Global Optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[12] D. R. Jones, M. J. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[13] M. D. McKay, R. J. Beckman, and W. J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, vol. 21, no. 2, pp. 239–245.

[14] M. Sasena, *Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations*. PhD thesis, University of Michigan, USA, 2002.

[15] D. R. Jones, C. D. Perttunen, and B. E. Stuckman, "Lipschitzian optimization without the Lipschitz constant," *Journal of Optimization Theory and Applications*, vol. 79, no. 1, pp. 157–181, 1993.

[16] F. Pukelsheim, "The Three Sigma Rule," *The American Statistician*, vol. 48, no. 2, 1994.

TABLE II: Hyperparameter spaces for the candidate methods

3-sigma	Student	GLR	SPRT	CUSUM	RSS
$\nu \in [0.5; 10]$	$N \in [50; 250]$	$N \in [10; 150]$ $\lambda \in [1; 10]$	$N \in [10; 150]$ $\mu_1 \in [0.1; 5]$ $\alpha \in [0.05; 0.2]$ $\beta \in [0.05; 0.2]$	$\delta \in [0.01; 5]$ $\lambda \in [0.1; 20]$	$N \in [10; 150]$ $q \in [5; 30]$ $M \in [200; 300]$

TABLE III: Gaussian test-case with criterion c_1

	3-sigma	Student	GLR	SPRT	CUSUM	RSS
Ranking	*6*	*4*	*5*	*3*	*1*	*2*
Best cost c_1	0.7573	0.0559	0.0699	0.0359	0.024	0.0339
t_{dt} (not optimized here)	501	28	35	18	12	17
r_{fd}	0.2124	0	0	0	0	0
r_{nd}	0.5449	0.0559	0.0699	0.0359	0.024	0.0339
Hyperparameter values	$\nu = 1.154$	$N = 216$	$N = 72$ $\lambda = 1.679$	$N = 24$ $\mu_1 = 0.5296$ $\alpha = 0.1017$ $\beta = 0.1862$	$\delta = 0.2872$ $\lambda = 4.8907$	$N = 38$ $q = 30$ $M = 250$

TABLE IV: Uniform test-case with criterion c_1

	3-sigma	Student	GLR	SPRT	CUSUM	RSS
Ranking	*6*	*5*	*4*	*3*	*1*	*2*
Best cost c_1	0.7585	0.1141	0.0978	0.0359	0.01	0.0339
t_{dt} (not optimized here)	501	37	47	285	5	30
r_{fd}	0.002	0.0741	0.004	0	0	0
r_{nd}	0.7565	0.0399	0.0938	0.0359	0.01	0.0339
Hyperparameter values	$\nu = 1.8304$	$N = 57$	$N = 55$ $\lambda = 2.6364$	$N = 20$ $\mu_1 = 0.6448$ $\alpha = 0.1248$ $\beta = 0.1895$	$\delta = 0.3762$ $\lambda = 5.5459$	$N = 20$ $q = 29$ $M = 300$

TABLE V: Gaussian test-case with criterion c_2

	3-sigma	Student	GLR	SPRT	CUSUM	RSS
Ranking	*6*	*4*	*5*	*3*	*1*	*2*
Best cost c_2	5.7691	0.321	0.4199	0.2159	0.144	0.1839
t_{dt}	501	26	35	18	12	15
r_{fd}	0.1583	0	0	0	0	0.004
r_{nd}	0.6008	0.061	0.0699	0.0359	0.024	0.0299
Hyperparameter values	$\nu = 1.3445$	$N = 217$	$N = 72$ $\lambda = 1.679$	$N = 24$ $\mu_1 = 0.5296$ $\alpha = 0.1017$ $\beta = 0.1862$	$\delta = 0.2872$ $\lambda = 4.8907$	$N = 39$ $q = 28$ $M = 299$

TABLE VI: Uniform test-case with criterion c_2

	3-sigma	Student	GLR	SPRT	CUSUM	RSS
Ranking	*6*	*5*	*4*	*3*	*1*	*2*
Best cost c_2	5.7685	0.4841	0.5318	0.2199	0.06	0.2079
t_{dt}	501	37	44	18	5	18
r_{fd}	0.7585	0.0741	0.004	0.004	0	0
r_{nd}	0.7565	0.0399	0.0878	0.0359	0.01	0.0359
Hyperparameter values	$\nu = 1.8478$	$N = 57$	$N = 52$ $\lambda = 2.4794$	$N = 25$ $\mu_1 = 0.5537$ $\alpha = 0.1583$ $\beta = 0.0639$	$\delta = 0.3543$ $\lambda = 5.6311$	$N = 21$ $q = 25$ $M = 300$

[17] W. S. Gosset, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.

[18] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[19] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall Englewood Cliffs, NJ, 1993.

[20] F. Gustafsson, *Adaptive Filtering and Change Detection*. Wiley London, 2001.

[21] E. Weyer, K. Sangho, and M. C. Campi, "A randomised subsampling method for change detection," in *Proceedings of the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, SAFEPROCESS 2009, Barcelona, Spain*, 2009.

[22] M. Bartyś, R. J. Patton, M. Syfert, S. de las Heras, and J. Quevedo, "Introduction to the DAMADICS actuator FDI benchmark study," *Control Engineering Practice*, vol. 14, no. 6, pp. 577–596, 2006.

[23] J. Villemonteix, E. Vazquez, and E. Walter, "An informational approach to the global optimization of expensive-to-evaluate functions," *Journal of Global Optimization*, vol. 44, no. 4, pp. 509–534, 2009.