A benchmark of kriging-based infill criteria for noisy optimization

Victor Picheny, Tobias Wagner, David Ginsbourger

Structural and Multidisciplinary Optimization manuscript No.
(will be inserted by the editor)

A benchmark of kriging-based infill criteria for noisy optimization

Victor Picheny · Tobias Wagner · David Ginsbourger

October 2011

Abstract Responses of many real-world problems can only be evaluated perturbed by noise. In order to make an efficient optimization of these problems possible, intelligent optimization strategies successfully coping with noisy evaluations are required. In this article, a comprehensive comparison of existing kriging-based methods for the optimization of noisy functions is provided. Ten methods are described using a unified formalism, and compared on analytical benchmark problems with different configurations (noise level, maximum number of observations, initial number of observations). It is found that the optimal method depends on the optimization problem, even though some criteria are consistently more efficient than others.

1 Introduction

The use of kriging for modeling and optimizing deterministic computer simulations has a long and successful tradition [23,14,29,15]. In recent years, there has been an increasing interest in the study of "stochastic" simulators or processes, whose outputs can only be observed in the presence of noise. Examples of such simulators can be found in a wide area of applications, including nuclear safety assessment [6], discrete event simulation [1], acoustic wave propagation in turbulent fluids [11], robust airfoil optimization [17], design of composite materials [25,24] and experimental measurements in mechanical engineering [3].

CERFACS (Centre Europeen de Recherche et de Formation Avancee en Calcul Scientifique), 42 avenue G. Coriolis, 31057 Toulouse, France. E-mail: picheny@cerfacs.fr · TU Dortmund, Baroper Straße 301, 44227 Dortmund, Germany · Bern University, Alpeneggstrasse 22, 3012 Bern, Switzerland

The large variety of applications has resulted in many different approaches for *Noisy Kriging-based Optimization* (NKO) over the last years [7,10,30,19]. Most of these NKO algorithms use the same formulation of the kriging model. Consequently, their differences are only based on different formulations of the criterion for selecting the next evaluation point(s) – the so-called infill criterion. The ideas behind these criteria range from a pure exploration of the design space to an intensive reevaluation of the currently best solution(s). Since these approaches have been developed within different disciplines, they have only been compared to state-of-the-art approaches in the respective fields. Based on implementation issues with respect to the multi-dimensional optimization of the criteria, often only one-dimensional examples are used. An interdisciplinary benchmark has not been performed until now.

In this paper, a comprehensive benchmark of the different NKO algorithms on test functions of varying dimension is thus conduced. The benchmark focuses on the class of problems shared by all these NKO approaches: a box-constrained, real-valued search space which is mapped to a single objective function $y : \mathbf{x} \in D \subset \mathbb{R}^d \rightarrow y(\mathbf{x}) \in \mathbb{R}$, where experiments can only provide noisy observations $\widetilde{y_i} = y(\mathbf{x}^i) + \epsilon_i$ $(1 \leq i \leq n)$ of the true response $y(\mathbf{x}^i)$. In order to stick to the assumptions made in most of the approaches, the observation noises ϵ_i are considered as being Gaussian, centered and i. i. d. This results in a performance analysis of the approaches under ideal conditions. Based on a set of test functions covering important problem properties, such as uni- and multi-modality, low and moderate search space dimensions, and different noise levels (variances), the relative strengths and drawbacks of the different NKO approaches are revealed.

Since all NKO approaches are based on a kriging model which is sequentially refined by new observations, they share important parameters, such as the size of the initial design of experiments and the choice of the covariance kernel. A systematic analysis of these parameters within the benchmark can thus assist in finding suitable settings and in identifying interactions between them and the corresponding NKO approach. This is the basis for a fair comparison of the NKO algorithms.

Before the design and the results of the benchmark are presented, the kriging model is described. Based on this background, the different infill criteria of the NKO algorithms are presented and formally compared. The implementation of the benchmark and solutions to some subproblems are explained in section 4.3. Finally, the experiments are described and the results are discussed. The paper is concluded with a summary of the results and an outlook on further research topics in NKO.

2 The Kriging model

Kriging is a functional approximation method originally coming from geosciences [16], and having been popularized in the computer experiments [23] and machine learning [21] communities. The basic idea behind kriging is the assumption that the response of interest y can be considered as one realization of a Gaussian process Y. In this work, we particularly focus on *Ordinary Kriging* (OK) which is described in the following.

2.1 Ordinary kriging

In the OK framework [18], $Y(\mathbf{x})$ is assumed to be of the form:

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}) \tag{1}$$

where $\mu \in \mathbb{R}$ is an unknown constant trend, and Z is a Gaussian process with zero mean and stationary (translation-invariant) covariance kernel of the form $k : (\mathbf{x}, \mathbf{x}') \in D^2 \to k(\mathbf{x}, \mathbf{x}') = \sigma^2 r(\mathbf{x} - \mathbf{x}'; \psi)$ for an admissible correlation function r with parameters ψ.

Under such hypotheses [1], the kriging model can be simply defined as the expectation and variance of Y conditionally on the observations:

$$m(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x})|Y(\mathbf{x}^i) = y_i, 1 \leq i \leq n] \tag{2}$$

$$s^2(\mathbf{x}) = \mathrm{Var}[Y(\mathbf{x})|Y(\mathbf{x}^i) = y_i, 1 \leq i \leq n] \tag{3}$$

where | means "conditional on".

$m(\mathbf{x})$ is called the kriging mean. It provides an interpolator for each observation $\mathbf{x}^i$ by enhancing the constant trend based on the correlation to the existing observations. $s^2(\mathbf{x})$ denotes the kriging variance (or prediction variance), which can be seen as a point-wise estimator of the model uncertainty. $m(\mathbf{x})$ and $s^2(\mathbf{x})$, after conditioning on n observations, are given by the following equations:

$$m_n(\mathbf{x}) = \widehat{\mu}_n + \mathbf{k}_n(\mathbf{x})^T \mathbf{K}_n^{-1}(\mathbf{y}^n - \widehat{\mu}_n \mathbf{1}_n), \tag{4}$$

$$s_n^2(\mathbf{x}) = \sigma^2 - \mathbf{k}_n(\mathbf{x})^T \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}) + \frac{\left(1 - \mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x})\right)^2}{\mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{1}_n}, \tag{5}$$

with:

- $\mathbf{y}^n = (y_1, \ldots, y_n)^T$,
- $\mathbf{K}_n = \left(k(\mathbf{x}^i, \mathbf{x}^j)\right)_{1 \leq i,j \leq n}$,
- $\mathbf{k}_n(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}^1), \ldots, k(\mathbf{x}, \mathbf{x}^n))^T$,
- $\mathbf{1}_n$ is a $n \times 1$ vector of ones, and
- $\widehat{\mu}_n = \mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{y}^n / \mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{1}_n$ is the best linear unbiased estimate of μ.

The kriging mean can be written as an (adaptively) weighted sum of the observations:

$$m_n(\mathbf{x}) = \boldsymbol{\lambda}^n(\mathbf{x}) \mathbf{y}^n, \tag{6}$$

with $\boldsymbol{\lambda}^n(\mathbf{x}) = \left(\mathbf{k}_n(\mathbf{x})^T + \frac{(1 - \mathbf{k}_n(\mathbf{x})^T \mathbf{K}_n^{-1} \mathbf{1}_n)}{\mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{1}_n} \mathbf{1}_n^T\right) \mathbf{K}_n^{-1}$. Kriging is thus often referred to as *best linear predictor*.

2.2 Kriging with noisy observations

In the framework of noisy observations, the $\widetilde{y}_i$ can be considered as realizations of the random variables $\widetilde{Y}_i := Y(\mathbf{x}^i) + \varepsilon_i$, so that Kriging amounts to conditioning Y on the noisy observations $\widetilde{Y}_i$ $(1 \leq i \leq n)$. As shown earlier in [8], provided that the process Y and the Gaussian measurement errors ε_i are stochastically independent, the process Y is still Gaussian conditionally on the noisy observations $\widetilde{Y}_i$ $(1 \leq i \leq n)$. Its conditional mean and variance functions are given by similar OK equations, with the only difference that $\mathbf{K}_n$ is replaced by $\widetilde{\mathbf{K}}_n = \mathbf{K}_n + \tau^2 \mathbf{I}_n$ at every occurrence in the m_n, s_n^2 and $\widehat{\mu}_n$ equations, where τ^2 is the known or estimated variance of the noise variables ε_i.

In geostatistics, as well as in computer experiments, an alternative, but equivalent, formulation $\widetilde{\mathbf{R}}_n = \frac{\widetilde{\mathbf{K}}_n}{\sigma^2 + \tau^2}$ is often used in order to write the kriging equations in terms of correlations. In this case, $\widetilde{R}_{ij} = 1$ if $i = j$, and $(1 - \nu) r(\mathbf{x}_i, \mathbf{x}_j)$ otherwise, where τ^2 is called *nugget* and $\nu = \frac{\tau^2}{\sigma^2 + \tau^2}$ *scaling factor* [27]. Another equivalent formulation, using variograms, can be found under the name of *ns-kriging* [24].

[1] Assuming further that μ is independent of Z and follows an improper uniform distribution over $\mathbb{R}$.

Note also that the model presented here resembles but slightly differs from the so-called *kriging with nugget effect* of the geostatistics literature [18], where the error variance τ^2 appears also in the covariance vector $\mathbf{k}_n(\mathbf{x})$, which makes it a discontinuous, interpolating model.

In the case of heterogeneous noise variances, i.e. when $var(\tilde{y}_1) = \tau_1^2 \neq \ldots \neq var(\tilde{y}_n) = \tau_n^2$, $\tau^2\mathbf{I}_n$ is replaced by $diag(\left[\tau_1^2 \ldots \tau_n^2\right])$. In our framework, the observation noise is homoskedastic, but a generalized model is used for the EQI criterion computation (see 3.5).

Contrarily to the noiseless case, $m_n(.)$ is not interpolating noisy measurements and $s_n^2(.)$ does not vanish at that points. Figure 1 shows an example of kriging based on noisy observations.

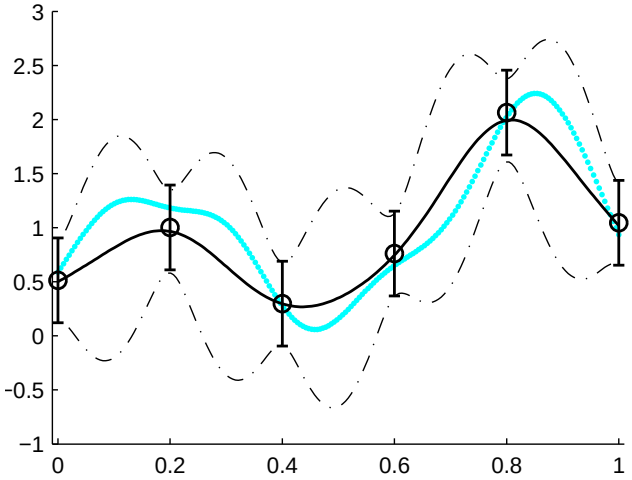


Fig. 1 Actual function (bold gray), Kriging mean (bold black) and 90% confidence intervals (mixed line); the circles are the observation values $\tilde{y}_i$, the bars show the noise amplitude ($\pm 2\tau$).

2.3 Covariance functions

A large variety of covariance kernels are available in the literature (see , e. g., [26] or [21] for a detailed summary). The choice of the kernel and the value of its parameters determine the shape (smoothness, amplitude of the prediction variance, ...) of the kriging model. In this work, two kernels are considered:

– the Gaussian anisotropic kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left[-\sum_{j=1}^{d}\left(\frac{x_j - x'_j}{\theta_j}\right)^2\right] \tag{7}$$

– the Matérn anisotropic kernel with $\nu = 3/2$:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2\left[1 + \sqrt{3}\sum_{j=1}^{d}\frac{\left|x_j - x'_j\right|}{\theta_j}\right] \times \exp\left[-\sum_{j=1}^{d}\frac{\left|x_j - x'_j\right|}{\theta_j}\right] \tag{8}$$

Both kernels depend on a set of parameters, σ^2 and $\{\theta_1, \ldots, \theta_d\}$, which are often referred to respectively as *process variance* and *ranges*.

2.4 Covariance parameter estimation

Covariance parameters are usually not known beforehand by the user and are estimated based on the observation vector $\tilde{\mathbf{y}}^n$. To accomplish this, several methods are available, e. g., maximum-likelihood-based approaches, (semi-)variograms, or cross-validation. Here, we focus on maximum-likelihood estimation (MLE).

σ^2 and the θ_i's are estimated by maximizing the probability density function of $\tilde{\mathbf{Y}}^n$ under the assumption of a multivariate Gaussian distribution:

$$L = (2\pi)^{-\frac{n}{2}}\det\left[\bar{\mathbf{K}}_n\right]^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\tilde{\mathbf{y}}^n - \widehat{\mu}_n\mathbf{1}_n\right)^T\bar{\mathbf{K}}_n^{-1}\left(\tilde{\mathbf{y}}^n - \widehat{\mu}_n\mathbf{1}_n\right)\right) \tag{9}$$

or equivalently by minimizing the log-likelihood (omitting constants):

$$l = \log\left(\det\left[\bar{\mathbf{K}}_n\right]\right) + \left(\tilde{\mathbf{y}}^n - \widehat{\mu}_n\mathbf{1}_n\right)^T\bar{\mathbf{K}}_n^{-1}\left(\tilde{\mathbf{y}}^n - \widehat{\mu}_n\mathbf{1}_n\right) \tag{10}$$

In the noiseless case, there exists an explicit expression for the optimal σ^2 as a function of the θ_i, which allows the problem to be simplified to the optimization of the θ_i (*concentrated log-likelihood*). Unfortunately, the superposed noise variance τ^2 prevents us from doing so here, so the optimization of equation 10 needs to be performed with respect to the whole vector of parameters:

$$[\hat{\sigma}^2, \hat{\theta}_1, \ldots, \hat{\theta}_n] = \arg\min l(\sigma^2, \theta_1, \ldots, \theta_n) \tag{11}$$

where the dependency of l on the parameters appears through $\bar{\mathbf{K}}_n$ and $\widehat{\mu}_n$. If τ^2 is an unknown quantity, it can also be considered as a parameter of the log-likelihood optimization.

3 Infill criteria

In the sequential procedure of most kriging-based optimizers, the design to be evaluated next is determined

based on the optimization of a so-called infill criterion. This infill criterion uses information of the current model in order to assess the utility of evaluating this design on the actual problem. In this section, we present definitions and ideas behind the infill criteria analyzed in our benchmark.

3.1 The classical Expected Improvement

The *Expected Improvement* (EI) has probably become the most popular infill sampling criterion for kriging-based global optimization of expensive-to-evaluate deterministic functions following the seminal paper of Jones et al. [14]. Let

$$I_n(\mathbf{x}) := \left(\min_{1 \leq i \leq n} (Y(\mathbf{x}^i)) - Y(\mathbf{x}) \right)^+ \tag{12}$$

denote the *improvement* obtained by evaluating Y at $\mathbf{x}$ after the n^{th} iteration, where $(.)^+ := \max(0, .)$. EI_n is defined as the expectation of I conditionally on the observations:

$$\begin{aligned} EI_n(\mathbf{x}) &= \mathbb{E}\left[I_n(\mathbf{x}) | Y(\mathbf{X}^n) = \mathbf{y}^n \right] \\ &= \mathbb{E}[(y_{min} - Y(\mathbf{x}))^+ | Y(\mathbf{X}^n) = \mathbf{y}^n] \end{aligned} \tag{13}$$

where $y_{min} := \min(y(\mathbf{X}^n))$ denotes the currently known minimum at the n^{th} iteration.

As shown in [14], the EI is fortunately analytically tractable

$$\begin{aligned} EI_n(\mathbf{x}) = (y_{min} - m_n(\mathbf{x}))\Phi\left(\frac{y_{min} - m_n(\mathbf{x})}{s_n(\mathbf{x})} \right) \\ + s_n(\mathbf{x})\phi\left(\frac{y_{min} - m_n(\mathbf{x})}{s_n(\mathbf{x})} \right), \end{aligned} \tag{14}$$

where Φ and ϕ denote the Gaussian cumulative distribution function and density function, respectively.

In the sequential procedure, the next measurement is performed where EI is maximum:

$$\mathbf{x}^{n+1} = \arg\max_{\mathbf{x} \in D} EI_n(\mathbf{x}). \tag{15}$$

By construction, EI_n is always non-negative, strictly increasing with s_n and decreasing with m_n[2]. Furthermore, it can be shown that for an interpolating kriging model, $\forall \mathbf{x} \in \mathbf{X}^n$, $EI_n(\mathbf{x}) = 0$ holds. Hence, maximizing EI_n never leads to re-evaluating y at already sampled points.

In the framework of noisy observations, EI will depart from this property since s_n is not necessarily 0 at $\mathbf{x} \in \mathbf{X}^n$. Moreover, the true minimum $\min(y(\mathbf{X}^n))$ at time n is not exactly known due to the noise on the observations.

[2] We consider minimization problems in this paper.

3.2 Expected improvement with "plug-in"

One possibility to deal with the fact that $\min_{1 \leq i \leq n}(y(\mathbf{x}^i))$ is not exactly known at time n is to replace it by some arbitrary target T, meant to be an efficient surrogate of y_{min}. This leads to the so-called *Expected Improvement with plugin*, denoted here by $EI_{T,n}$:

$$EI_{T,n}(\mathbf{x}) = \mathbb{E}[(T - Y(\mathbf{x}))^+] \tag{16}$$

The choice of T is an important issue, since too high or too low values may have a significant influence on the shape of $EI_{T,n}$ and thus change its behavior relatively to EI with known y_{min}. A first "naive" approach consists in choosing $T = \min\left(\tilde{y}^i\right)$, but this plugin lacks robustness since it suffices to have one noisy observation with a coincidentally low value to severely underestimate y_{min} for the rest of the optimization. Following the approach mentioned in [30], $T = \min\left(m_n(\mathbf{X}^n)\right)$ seems a sensible option. A generalization considered here is to take the minimum of Kriging β-quantiles at $\mathbf{X}^n$, for a level $\beta \in]0, 1[$ tuned by the user.

Whatever the chosen value for T, a nice fact about $EI_{T,n}$ is that it can be analytically calculated, as well as its gradient, just as the classical EI:

$$\begin{aligned} EI_{T,n}(\mathbf{x}) = (T - m_n(\mathbf{x}))\Phi\left(\frac{T - m_n(\mathbf{x})}{s_n(\mathbf{x})} \right) \\ + s_n(\mathbf{x})\phi\left(\frac{T - m_n(\mathbf{x})}{s_n(\mathbf{x})} \right). \end{aligned} \tag{17}$$

However, one drawback of $EI_{T,n}$ for noisy optimization is that it does not take into account the noise of the future observation: everything is calculated as if the next evaluation would be deterministic. The AEI criterion presented in the next section addresses this issue by adding a multiplicative term to $EI_{T,n}$, penalizing the points whose kriging variance s^2 is small compared to the noise level τ.

3.3 Augmented Expected Improvement (AEI)

The AEI criterion was proposed by Huang et al. ([10] for the noisy framework and [9] for multi-fidelity). The idea of surrogating the unknown y_{min} by the value of the Kriging mean at some point is also used. But this time, instead of considering $T = \min\left(m_n(\mathbf{X}^n)\right)$, T is taken as $m_n(\mathbf{x}^{**})$, where the so-called *effective best solution* $\mathbf{x}^{**}$ is obtained by minimizing $m_n + \alpha s_n$ over the already observed points in order to have a more robust estimate of the plugin. In other words, T is the Kriging mean value at the design point with lower β-quantile, where $\Phi(\beta) = \alpha$. The value $\alpha = 1$ is recommended by the authors.

Additionally, a multiplicative penalty is introduced in order to account for the noise variance of the next evaluation:

$$AEI_n(\mathbf{x}) = EI_{T,n}(\mathbf{x}) \times \left(1 - \frac{\tau}{\sqrt{s_n^2(\mathbf{x}) + \tau^2}}\right), \qquad (18)$$

The penalty term is one if $\tau = 0$, and decreases towards zero when τ increases. It thus reduces to the original EI function whenever $\tau = 0$. Huang et al. justify the penalty to "account for the diminishing return of additional replicates as the predictions become more accurate". In fact, it penalizes designs with small prediction variance $s_n^2(\mathbf{x})$ and therefore enhances exploration.

3.4 The reinterpolation procedure

The *reinterpolation* method was proposed by Forrester et al. [7]. Instead of modifying the EI criterion for the noisy case, the authors propose to use simultaneously a kriging with noisy observations (as defined in equations 4 and 5) -called the *regressing* model- and an *interpolating* kriging, which is built as follows: The covariance structure and its parameters, as well as the DOE, of the regressing model are inherited, but the kriging mean predictions of the regressing kriging at the DOE points are used as observation vector. Since this latter model is noise-free, the classical EI can be used as an infill criterion. Summarizing, the reinterpolation procedure consists of four steps:

1. Build a kriging based on the noisy observations $\widetilde{\mathbf{y}}^n$
2. Compute the kriging predictor at the DOE points $m_n(\mathbf{x}^1), \ldots, m_n(\mathbf{x}^n)$
3. Build an interpolating kriging model using $\mathbf{X}^n$ and $\mathbf{y}^n = [m_n(\mathbf{x}^1), \ldots, m_n(\mathbf{x}^n)]^T$
4. Solve $\mathbf{x}^* = argmax EI_n(\mathbf{x})$ using the interpolating model.

Note that this procedure was initially designed for "deterministic" noise, due to numerical instabilities and ill-posedness of the simulated system. In that case, two very close designs would return different results, but repeating the same experiment would return the same output. Hence, the reinterpolating procedure does not allow repetitions – EI is zero at already observed designs.

3.5 Expected Quantile Improvement (EQI)

The main concept of the EQI criterion, as detailed in [19], is that in a noisy situation, improvement may be measured in the model rather than on the noisy data, the measure of reference being the kriging quantile $q_n(\mathbf{x}) = m_n(\mathbf{x}) + \Phi(\beta)s_n(\mathbf{x})$ (with $\beta \in [0.5, 1[$) rather than the noisy observations. Similarly to the noiseless case, an improvement between steps n and $n+1$ is defined as:

$$I_n(\mathbf{x}) := \left(\min_{1 \le i \le n}(q_n(\mathbf{x}^i)) - Q_{n+1}(\mathbf{x})\right)^+ \qquad (19)$$

where Q_{n+1} is the quantile of the kriging updated with $\mathbf{x}^{n+1} = \mathbf{x}$. EQI is defined as $\mathbb{E}[I_n(\mathbf{x})]$. It has been shown that the distribution of $Q_{n+1}(\mathbf{x})$ conditionally on the observations is Gaussian and analytically derivable, which leads to the following formula for the EQI:

$$\begin{aligned} EQI_n(\mathbf{x}) = (q_{min} - m_Q(\mathbf{x}))\Phi\left(\frac{q_{min} - m_Q(\mathbf{x})}{s_Q(\mathbf{x})}\right) \\ + s_Q(\mathbf{x})\phi\left(\frac{q_{min} - m_Q(\mathbf{x})}{s_Q(\mathbf{x})}\right), \end{aligned} \qquad (20)$$

where $q_{min} := \min_{1 \le i \le n}(q_n(\mathbf{x}^i))$ is the current best quantile and m_Q and s_Q denote the mean and standard deviation of the future quantile $Q_{n+1}(\mathbf{x})$, respectively.

In practice, this method requires to build, for each candidate $\mathbf{x}$, a kriging augmented with the observation $\mathbf{x}^{n+1} = \mathbf{x}$, $\widetilde{y}^{n+1} = m_n(\mathbf{x})$ and $\tau_{n+1}^2 = \tau_{new}^2$ (the noise of the future observation). The quantile mean and variance can then be simply extracted using:

$$m_Q(\mathbf{x}) = m_{n+1}(\mathbf{x}) + \phi^{-1}(\beta)s_{n+1}(\mathbf{x}) \qquad (21)$$

$$s_Q^2(\mathbf{x}) = \left(\lambda_{n+1}^{n+1}(\mathbf{x})\right)^2 \left(s_n^2(\mathbf{x}) + \tau_{new}^2\right) \qquad (22)$$

with λ_{n+1}^{n+1} being the $(n+1)$th term of the weight vector: $\boldsymbol{\lambda}^{n+1}(\mathbf{x}) =$

$$\left(\mathbf{k}_{n+1}(\mathbf{x})^T + \frac{(1 - \mathbf{k}_{n+1}(\mathbf{x})^T \mathbf{K}_{n+1}^{-1} \mathbf{1}_{n+1})}{\mathbf{1}_{n+1}^T \mathbf{K}_{n+1}^{-1} \mathbf{1}_{n+1}} \mathbf{1}_{n+1}^T\right) \mathbf{K}_{n+1}^{-1}$$

The future noise τ_{new}^2 accounts for the limited optimization budget, and is set to $\tau^2/(N-n)$, where N is the maximum number of observations. It is thus assumed that the remaining budget is completely spent for this solution, which is actually not desired. The above-defined rule can be seen as a heuristic in order to slightly shift the focus of the optimization from exploration to exploitation.

3.6 Alternatives to EI-based criteria

The last method considered in this benchmark is perhaps the most natural of the metamodel-based procedures and acts as a baseline for the other criteria. It consists of performing the next measurement where the current kriging mean or quantile is minimum:

$$\mathbf{x}^{n+1} = argmax_{\mathbf{x} \in D} m_n(\mathbf{x}) + \alpha \times s_n(\mathbf{x}) \qquad (23)$$

Although recognized as less efficient compared to the EI in the case of deterministic experiments [13], this

method seems worth studying in this benchmark because it is the only method whose behavior is independent of the presence of noise. It also has shown successful applications in kriging-based multi-objective optimization [5,20].

4 Design of the benchmark

4.1 Analytical test functions

As test problems, we employed six widely used analytical benchmark problems [4]. Their definitions are given in table 1. The original functions have been rescaled to map their search space to $[0,1]^d$, their mean to zero, and their variance to one. For the separable sphere function, the input vectors are shifted and rotated before evaluation. These functions being deterministic, the observation noise is added artificially using i. i. d. Gaussian random variables. The noise variance is chosen as explained in section 4.2.

The test functions are chosen to cover a large variety of problem properties. *Rosenbrock4* and *Sphere6* are unimodal functions. The valley of the global minimum is easy to find, however fine convergence to the global minimum is difficult. Since *Rosenbrock4* and *Sphere6* have a very low activity, the range covariance bounds are chosen as $[0.5, 5]$, which allows to have a very "flat" kriging model. *Branin-Hoo, Goldstein-Price, Hartman4* and *Hartman6* are multimodal functions. Here, the range bounds are set to $[0.1, 1]$, which allows to model high activity responses.

4.2 Algorithmic factors

A large number of factors can influence the quality of the different kriging-based procedures, whereby two types of factors can be distinguished:

- The factors related to the parameterization of the problem or optimization task.
- The factors for setting up the approach (usually tuned by the user).

In this benchmark, we consider the problem factors which we expect to have a significant influence on the performance of the different criteria. These factors are the modality of the problem, the dimension of the search space d, the noise level, and the allowed budget of evaluations (a similar classification can be found in [12]). For the tuning factors, we selected the ones which have been changed within different studies [2,3], i. e., the proportion of observations for the initial DOE and the choice of the covariance kernel.

Table 2 Summary of the benchmark factors and levels.

Factor	**Values**
Modality	uni- (*Rosenbrock4*, *Sphere6*) and multimodal (*Branin*, *GoldsteinPrice*, *Hartman4*, *Hartman6*)
Search space dimension d	2 (*Branin*, *GoldsteinPrice*), 4 (*Hartman4*, *Rosenbrock4*) and 6 (*Hartman6*, *Sphere6*)
Noise SD	5%, 20%, 50% (of the objective function SD)
Maximum number of evaluations	$20 \times d$, $40 \times d$
Number of initial evaluations	$4 \times d$, $10 \times d$
Covariance kernel	matern3/2, Gauss

For each factor to be considered, we have chosen two to three different values, as listed in Table 4.2. The noise level is expressed in terms of the proportion of the function standard deviation (SD) (which is one for all functions). The noise levels vary between moderate (5%) to extremely noisy (50%). In addition, for a given setup (including the infill criterion), result can depend on the initial DOE and on the noise realizations. To account for this variability, for each configuration 40 runs are performed with different initial DOEs and random seeds.

The total number of optimization runs performed for the benchmark is $n_{fct} \times n_{noises} \times n_{criteria} \times n_{covariances} \times n_{budgets} \times n_{DOEsizes} \times n_{runs}$=30,000.

The choice of the design type of the initial DOE is also probably a significant factor; however, here we fix it to be an LHS design optimized with respect to the maximin criterion, which is common practice in the kriging community. Other factors of minor importance, not considered here, may include the use of replications on the initial design [2], the choice of the method for covariance parameters estimation, the re-estimation or not of the covariance parameters during optimization, or the choice of the kriging trend (for a *Universal Kriging* model).

The reinterpolation procedure does not depend on any parameter; the AEI criterion depends on the penalization level α for the choice of the effective best solution, and is set to 1, as recommended by Huang et al. The EI with plugin of a quantile, EQI, and the quantile minimization depend on the quantile level β. For those methods, two levels ($\beta = 0.5$ and $\beta = 0.9$) are tested. A random search is performed as a baseline for the optimization performance. Table 3 summarizes the criteria and parameters tested in the benchmark.

In order to minimize the external variance in the comparison of the problem and tuning factors, the initial DOEs and observations have been re-used as much as possible. For instance, the same LHS is used for all

Table 1 Test functions.

Branin-Hoo (2D)	$y(\mathbf{x}) = \frac{1}{51.95}\left[\left(\bar{x}_2 - \frac{5.1\bar{x}_1^2}{4\pi^2} + \frac{5\bar{x}_1}{\pi} - 6\right)^2 + \left(10 - \frac{10}{8\pi}\right)\cos(\bar{x}_1) - 44.81\right]$ with: $\bar{x}_1 = 15 \times x_1 - 5$, $\bar{x}_2 = 15 \times x_2$
Goldstein-Price (2D)	$y(\mathbf{x}) = \frac{1}{2.427}\left[\log\left[\left(1 + (\bar{x}_1 + \bar{x}_2 + 1)^2(19 - 14\bar{x}_1 + 3\bar{x}_1^2 - 14\bar{x}_2 + 6\bar{x}_1\bar{x}_2 + 3\bar{x}_2^2)\right)\left(30 + (2\bar{x}_1 - 3\bar{x}_2)^2\,(18 - 32\bar{x}_1 + 12\bar{x}_1^2 + 48\bar{x}_2 - 36\bar{x}_1\bar{x}_2 + 27\bar{x}_2^2)\right)\right] - 8.693\right]$ with: $\bar{\mathbf{x}} = 4 \times \mathbf{x} - 2$
Rosenbrock4 (4D)	$y(\mathbf{x}) = \frac{1}{3.755 \times 10^5}\left[\sum_{j=1}^{3}\left(100(\bar{x}_{j+1} - \bar{x}_j^2)^2 + (1 - \bar{x}_i)^2\right) - 3.827 \times 10^5\right]$ with: $\bar{\mathbf{x}} = 15 \times \mathbf{x} - 5$
Hartman4 (4D)	$y(\mathbf{x}) = \frac{1}{0.839}\left[1.1 - \sum_{i=1}^{4} C_i \exp\left(-\sum_{j=1}^{4} a_{ji}\,(x_j - p_{ji})^2\right)\right]$ with: $\mathbf{C} = [1.0, 1.2, 3.0, 3.2]$ $\mathbf{a} = \begin{bmatrix} 10.00 & 0.05 & 3.00 & 17.00 \\ 3.00 & 10.00 & 3.50 & 8.00 \\ 17.00 & 17.00 & 1.70 & 0.05 \\ 3.50 & 0.10 & 10.00 & 10.00 \\ 1.70 & 8.00 & 17.00 & 0.10 \\ 8.00 & 14.00 & 8.00 & 14.00 \end{bmatrix}$, $\mathbf{p} = \begin{bmatrix} 0.1312 & 0.2329 & 0.2348 & 0.4047 \\ 0.1696 & 0.4135 & 0.1451 & 0.8828 \\ 0.5569 & 0.8307 & 0.3522 & 0.8732 \\ 0.0124 & 0.3736 & 0.2883 & 0.5743 \\ 0.8283 & 0.1004 & 0.3047 & 0.1091 \\ 0.5886 & 0.9991 & 0.6650 & 0.0381 \end{bmatrix}$
Hartman6 (6D)	$y(\mathbf{x}) = \frac{-1}{1.94}\left[2.58 + \sum_{i=1}^{4} C_i \exp\left(-\sum_{j=1}^{6} a_{ji}\,(x_j - p_{ji})^2\right)\right]$
Sphere6 (6D)	$y(\mathbf{x}) = \frac{1}{899}\left[\sum_{j=1}^{6} x_j^2 \times 2^j - 1745\right]$

Table 3 Summary of the infill criteria.

Criterion	**Parameter**	Abbreviation
Random search	-	RS
Reinterpolation	-	RI
AEI	-	AEI
EQI	$\beta = 0.5$	EQ50
	$\beta = 0.9$	EQ90
EI with plugin	$T = \min(\tilde{y}^i)$	PIy
	$T = \min(m_n(\mathbf{X}^n))$	PI50
	$T = \min(q_n(\mathbf{X}^n))$ with $\beta = 0.9$	PI90
Quantile minimization	$\beta = 0.5$	MQ50
	$\beta = 0.1$	MQ10

the test functions of the same dimension. The same LHS is used to generate the initial observations for the four different noise levels. The same set of initial observations is used for all the infill criteria. By these means, the LHSs can be used for grouping in nonparametric tests on the significance of the factor effects.

4.3 Implementation issues and solutions

4.3.1 Optimization of the kriging parameters

In all kriging-based procedures, providing accurate covariance parameters is a crucial point. In particular, the *range* parameters (θ_j in eqs. 7 and 8) reflect the predicted activity (or smoothness) of the objective function, which have a great effect on the shape of the infill criteria.

The parameter estimation is here done by maximum likelihood, as defined in section 2.4, using the R package *DiceKriging* [22]. Since the likelihood is known to often have local maxima for values corresponding to either very small range (white noise) or very large range (constant response), the covariance parameters are bounded to sensible intervals (see section 4.1). These intervals have been found by performing pre-experiments on the chosen test functions and are wide enough to cover the requirements of the different criteria.

In our setup, the parameters are estimated first using the initial DOE, and re-estimated after each additional measurement. The old parameters are included as potential candidates for the likelihood optimization, so the new parameters cannot be worse (in terms of likelihood) than the old ones. Nevertheless, it has been found that for some criteria, the parameter re-estimation sometimes fails due to numerical instability in the inversion of the covariance matrix. When this occurs, the model is updated based the old covariance parameters.

The reinterpolation technique is particularly sensitive to these problems since it uses an interpolating kriging on smoothed data. In case of failure, a small nugget is added to the interpolating model (in order to ease the covariance matrix inversion). If the model computation is still not possible, the run is terminated and the results for the last iteration are used.

4.3.2 Optimization of the infill criteria

At each optimization step, the infill criterion is maximized over D in order to choose the next measurement. This task can often become challenging, since the Expected Improvement and its "noisy" variants are known to be highly multimodal with some large "flat" regions (where the criterion takes values below the machine accuracy). Figure 2 shows an example of contour lines of the AEI criterion that illustrates these properties. Although the criteria are relatively inexpensive to compute (about 5 milliseconds for the EQI based on a kriging model with 50 points on a 2.9 GHz processor), an exhaustive search on a grid is not possible in dimensions higher than two because this optimization is performed in each iteration.

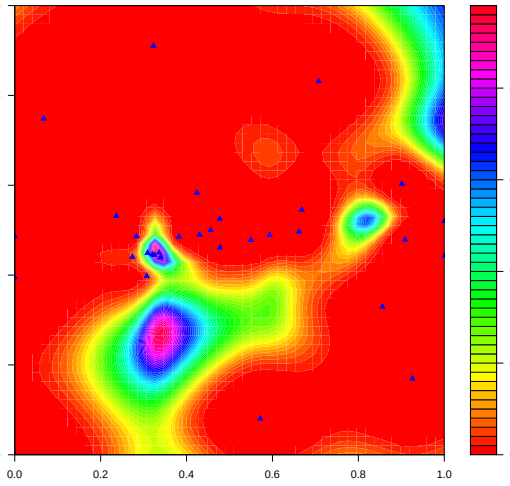


Fig. 2 Contour lines of the AEI during a typical optimization run (Goldsteinprice function, 29 points (triangles), noise level 20%).

Here, we chose to optimize the infill criteria using the *genoud* algorithm (GENetic Optimization Using Derivatives, [28]), which implements a hybrid of evolutionary algorithms and gradient descent. This algorithm allows the local optima to be accurately found thanks to the gradient descent while still having a good exploration of the search space due to the evolutionary algorithm. The analytical gradients of all the criteria for the ordinary kriging model have been calculated and implemented into *DiceKriging* [22].

To account for the increase of the complexity of the optimization with increasing search space dimension while ensuring a reasonable computational effort, the genoud parameters have been set to the following values:

- The population size is 6×2^d.
- The number of population generations is 20.
- The maximum number of evaluations within a gradient descent is 6×2^d.

With that setup, the global optima of the criteria are found in the vast majority of the cases for all considered configurations.

4.4 Research questions

The research questions addressed in the benchmark are directly related to the effects and interactions of the experimental and algorithmic factors varied in the benchmark. With respect to the algorithmic factors, it is of practical interest to know whether there is a specific best over all considered test instances. If this is not the case, the interactions between the factors of the problem instance and the algorithmic factor become important in order to assist in choosing the right NKO approach for a specific problem instance.

Based on the algorithmic factors considered in the benchmark, the main questions are: Is it possible to choose an optimal

1. covariance kernel,
2. size of the initial design, and
3. infill criterion

with respect to all considered

1. maximum budgets of evaluations,
2. noise levels of the test functions,
3. modalities of the test functions, and
4. decision space dimensions of the test functions

or are there specific choices depending on the level of the latter.

The above-mentioned research questions are particularly interesting for a practical user. For the research on NKO, also the effect of using specific information in the infill criteria is of interest. In this paper, we particularly focus on the effects of

1. the consideration of the current accuracy of an observation (e. g., by using a Kriging quantile rather than the Kriging mean)
2. the use of replications (Forrester's approach does not perform any replications while other infill criteria may do),
3. the complex computation of the EQI in comparison to simpler approximations, and
4. the consideration of the remaining evaluations (as in EQI),

For the first two questions, particularly the interaction with respect to the noise level of the test function is of interest. The last two questions are worth consideration in order justify or possibly decrease the complexity of the EQI approach.

5 Results

5.1 Observations

We start by analyzing the results quantitatively using the true objective value of the design which is identified as the current best by the corresponding infill criterion. In more detail, for RS and PIy we take the design point with the best noisy observation, for RI, PI50, EQ50, and MQ50, we take the design point with the best kriging mean, and for AEI, PI90 and EQ90 we take the one with the best kriging quantile.

5.1.1 A first sensitivity analysis

The first step of the analysis is to state whereas all the parameters considered in this benchmark had a significant influence on the algorithms performances. The structure of the benchmark makes it possible to analyze the effect of an isolated factor: for instance, for a fixed test function, noise level, budget, infill criterion and initial DOE size, the same LHS and initial set of observations are used with the Gaussian and the Matérn covariances, which allows a fair comparison between the two kernels (even though their respective performances depend on the noise measurements values during optimization). Since 40 LHS are used, we can compute and visualize for each configuration the proportion of runs for which one parameter value is better than the other. With $n = 40$ data, the 95% confidence interval for the proportion 0.5 is $[30\%, 70\%]$. The runs are compared based on the actual value of the function at the returned best solution.

According to the research questions, we first compare the results for the two covariance functions (Figure 3). The total number of configurations is $n_{fct} \times n_{noises} \times n_{criteria} \times n_{DOEsizes} \times n_{budgets} = 576$. We found 48 critical configurations (8%) outside the confidence interval of the proportion, 25 showing a better performance of the Gauss kernel over Matern and 23 the inverse. 10 configurations with the Rosenbrock function and 5% noise and 10 configurations with the Goldsteinprice function and the reinterpolation criterion that showed a better performance of the Gaussian kernel. No clear pattern appears for the other 28 critical configurations.

We can therefore conclude that the choice of the kernel is not a critical one. In some scenarios, however, the Gaussian kernel seems to be beneficial. We hence consider only the Gaussian kernel in the following analysis.

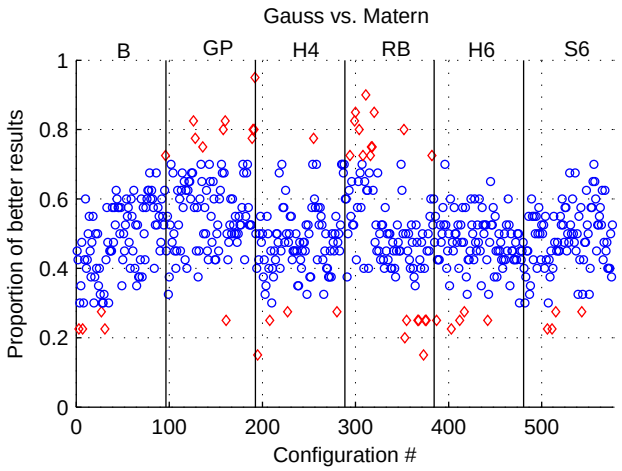


Fig. 3 Proportion of better performance of Gauss kernel against Matern kernel for all configurations. Proportions significantly higher than 50% are represented by red diamonds.

Then, we look at the influence of the DOE size. The total number of configurations is $n_{fct} \times n_{noises} \times n_{criteria} \times n_{covariances} \times n_{budgets} = 576$. All the experimental proportions are represented in Figure 4. Only 21 configurations (less that 3%) showed a significant difference, eight of them being for the Rosenbrock function with 5% noise and showed a better performance of the large DOE size. We can conclude that, for our benchmark, the DOE size has little influence on the optimization results. In the rest of the analysis, the large DOE size only is considered in order to reduce the number of configurations.

5.1.2 Comparison of criteria of the same family

Before we start the complete analysis of the infill criteria, we want to reduce the number of considered criteria in order avoid comparing with failed parameterizations and to improve the readability of the plots. To accomplish this, criteria of the same family are compared first. In particular, we compare MQ50 with MQ10, PI50 with PIy and PI90, and EQ50 and EQ90. The number of different configurations is here $n_{fct} \times n_{noises} \times n_{budgets} = 36$.

When comparing MQ10 and MQ50 (figure 5), 15 configurations showed no significant difference. For the other 21, MQ10 outperformed MQ50. So we can con-

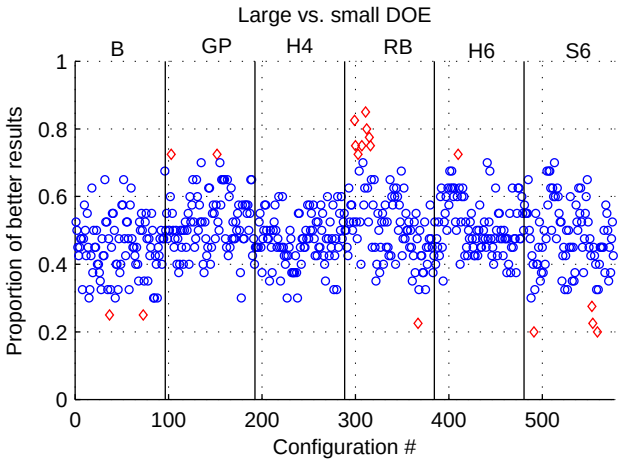


Fig. 4 Proportion of better performance of large DOE size against small DOE size for all configurations.

clude here that minimizing a low quantile is clearly a better option than minimizing the best predictor.

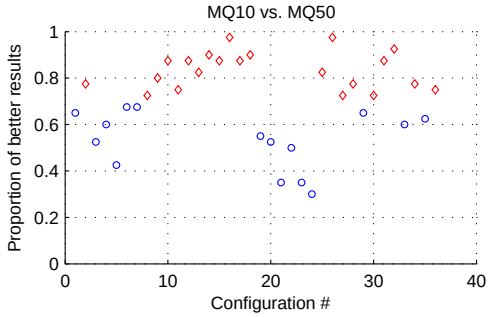


Fig. 5 Proportion of better performance of MQ10 against MQ50 for all configurations.

The results for PI50 and PI90 are shown in Figure 6. Here, 8 configurations (22%) were critical and all showed a better performance of PI50. Besides, although it is not individually statistically significant, most of the proportions are above 50%, which seems to indicate that PI50 is tendencially equal or better than PI90. We can then conclude that the plugin of the kriging mean outperforms the plugin of a high kriging quantile.

The other tests performed with criteria of the same family (PI50 and PI90 versus PIy, EQ50 versus EQ90, not shown here) did not show a clear outperformance of one method against another. The performance of PIy may lay between the ones of PI50 and PI90 since it is not significantly different from either of them.

5.1.3 Global rank analysis

Based on the previous observations, we limit now our analysis to the following criteria: RI, AEI, EQ50, EQ90,

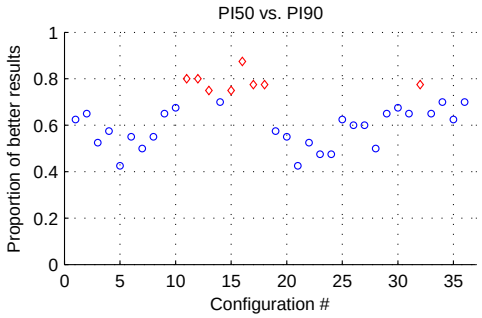


Fig. 6 Proportion of better performance of PI50 against PI90 for all configurations.

PI50, PIy and MQ10, and consider only the Gaussian kernel and the large DOE size. The random search (RS) results are given as reference value. For each configuration, we compute the average ranks of the criteria over the 40 LHS for all test functions, noise levels and budgets (36 configurations total).

We can see first that no method outperforms the others for all the configurations. However, we can observe that:

- AEI works best on Hartman4 for both budgets, and on Goldstein-Price and Hartman6 for the high budget
- RI works best on Branin and Sphere for the low budget and moderate noise levels
- MQ10 is best on Hartman6 with low budget
- EQ90 is best on Rosenbrock for the high noise level

Inversely, some methods are significantly less efficient on some configurations:

- EQ90 works poorly on Goldstein-Price with high noise and Hartman4
- PIy works poorly on Branin with high noise, Rosenbrock and Hartman6

Almost all the methods are significantly better than random search; however, on Branin with 50% noise, no method outperforms random search, and on Rosenbrock with low budget and for all noises, only PI50 and EQ90 are significantly better.

5.1.4 Detailed analysis using boxplots

In this section, we propose a detailed comparison of the methods. We represent the results in the form of boxplots of actual function values at best design for each method (figure 8). To limit the amount of figures, we show the results for the low budget only.

With 5% noise, all the criteria return almost similar results: accurate identification of the minimum for all

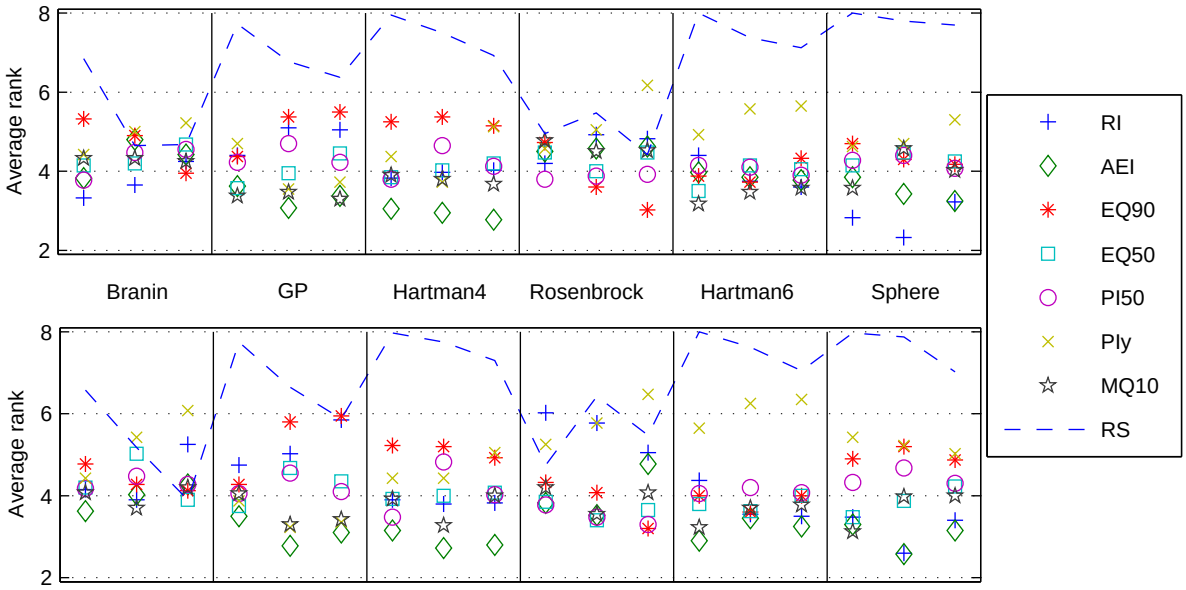


Fig. 7 Average ranks over the 40 LHS of the criteria for all test functions, noise levels and budgets. Upper figure: low budget; lower figure: large budget. For each box, the three columns correspond (from left to right) to the noise levels 5%, 20% and 50%.

runs for Branin (A), Hartman6 (E) and Sphere (F), accurate identification for most of the runs but a few outliers (runs trapped in local optima) for Goldstein-Price (B) and Hartman4 (C), approximate identification only for Rosenbrock (D).

PIy appears as very efficient on Goldstein-Price and Hartman4 with 20% noise, but also the worst method on Rosenbrock, Hartman6, Sphere. With 5% noise, EQ90 sometimes show large tails (Goldstein-Price, Hartman4, Sphere). AEI outperforms the other methods on Hartman4, both in terms of median and 95th percentile.

For the other configurations, the results in terms of either median of 95th percentile are less contrasted.

5.2 Discussion

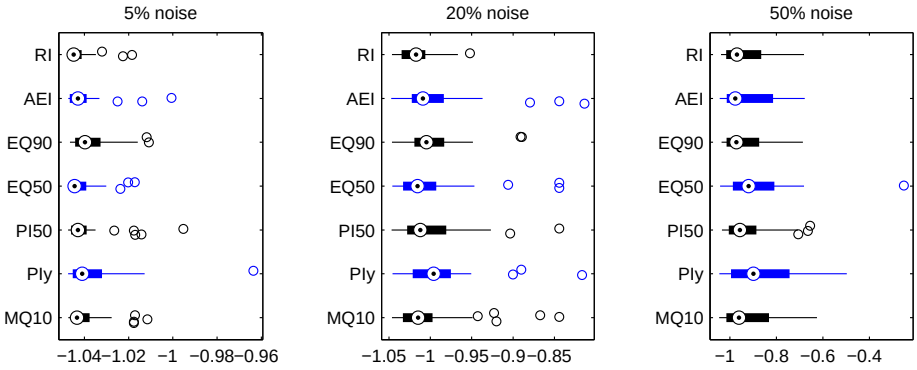
The first conclusion of this benchmark analysis is the limited influence of sample size on the optimization results compared to other parameters. Using smaller initial DOEs results in more optimization steps, which seems intuitively more efficient. However, using larger initial DOEs ensures a good initial exploration, which reduces the risk of converging to a local optimum, and tends to produce more accurate models. These effects seem to balance each other regarding the optimization efficiency.

Another parameter of limited influence is the choice of the covariance kernel, which is surprising since the two kernels considered here imply very different assumptions on the shape of the objective function (C^1 for matern 3/2, C^∞ for gauss). Here it seems to be dominated by other factors, such as the ability of kriging to fit non-stationary functions or the robustness of the covariance parameter estimation.

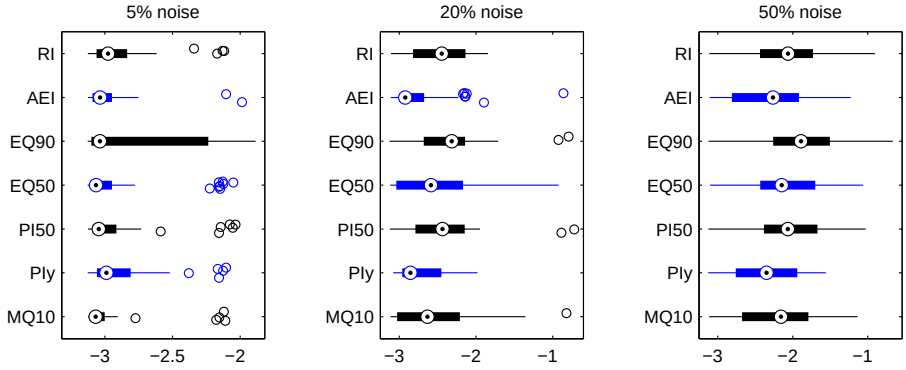
Out of the 10 criteria tested here, three can be considered as poor alternatives: $PI90$, PIy and $MQ50$. The criterion $MQ50$ was proposed essentially because it is relatively common practice in surrogate modeling to sequentially sampling at the minimum of the best predictor. Although known as a bad solution for deterministic functions [13], the question was left open in the noisy case. It is found that the $MQ50$ performances are also poor in noise so this solution is not competitive with other criteria, which is logical since it does not offer any trade-off between exploration and exploitation. The low quantile criterion $MQ10$, however, proved in our context to be competitive.

The poor performances of $PI90$ and PIy can be explained by looking at the EI equation 17. For $PI90$, by plugin a high quantile for T, the quantity $T - m_n$ is likely to be positive and large: we indeed replace y_{min} by a (very) positively biased estimate, which make the existing points look more interesting that they actually are and hinders exploration. With PIy, we also use a biased estimate ($min(\tilde{\mathbf{y}})$) of y_{min}. With high noise in particular, y_{min} is likely to be strongly underestimated, which results in increased exploration. Forcing exploration seems beneficial in some cases (on the Goldstein-Price function, for which the minimum is indeed in a

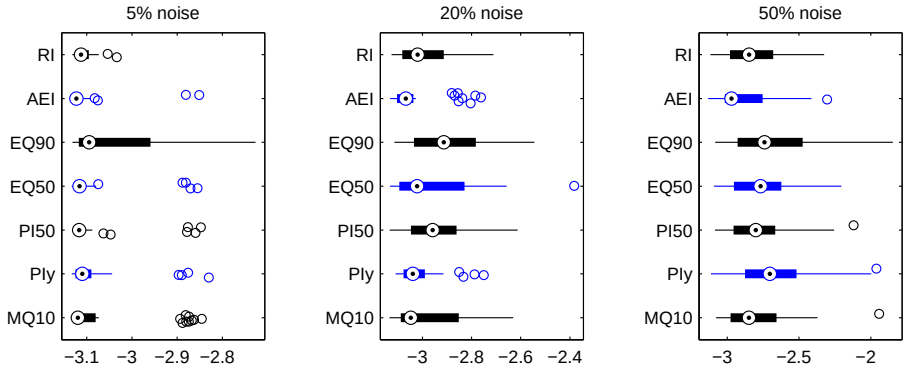
A) Branin:



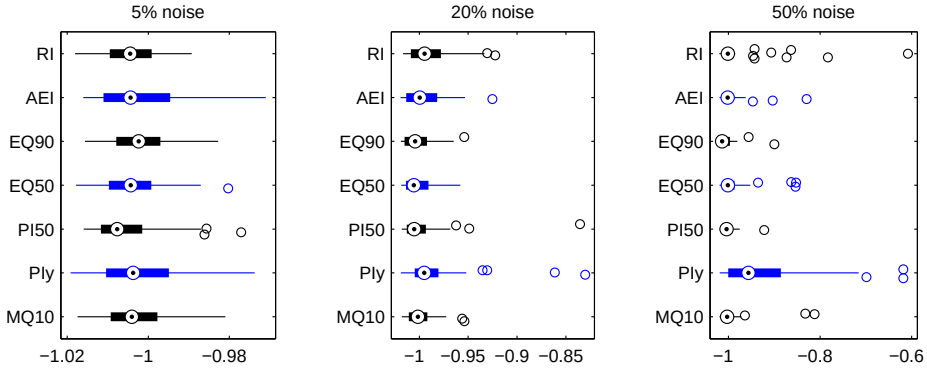
B) Goldstein-Price:



C) Hartman4:



D) Rosenbrock:



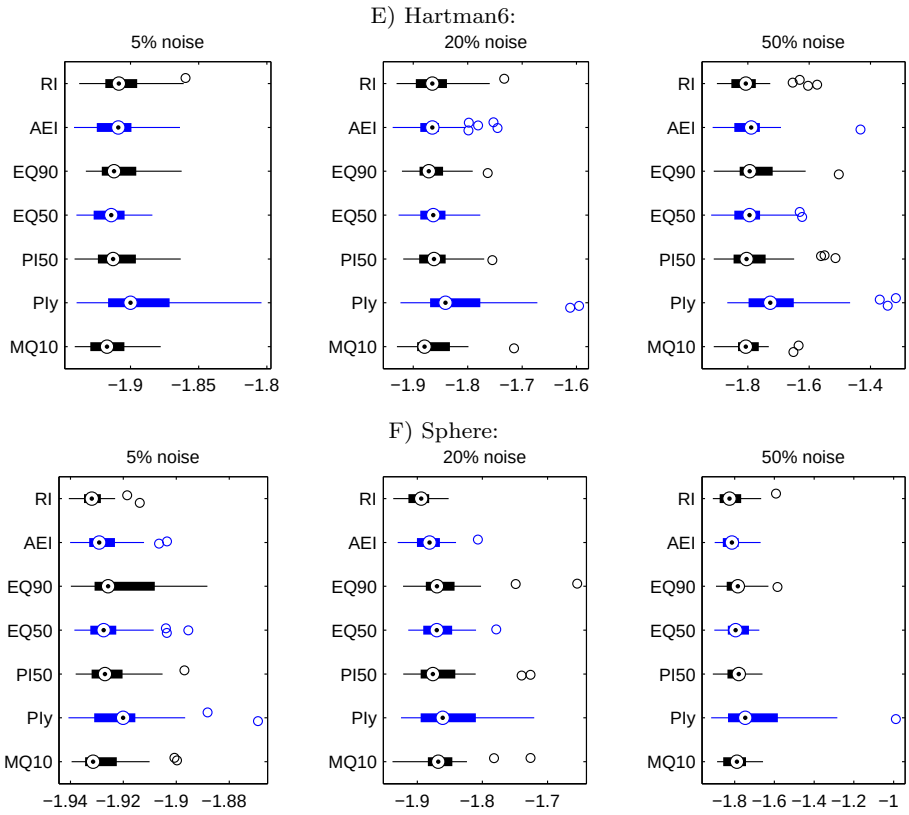


Fig. 8 Boxplots of actual function value at best design for each method.

small valley), but overall the criterion $PI50$ appears a better alternative.

The RI and $EQ90$ criteria show contrasted performances depending on the configurations. By construction, the RI criterion is quite exploratory (in particular, it does not allow replications), which can be beneficial for optimization and eases the covariance parameter estimation step, which explains the very good performances in some cases. However, one can observe that the RI performances decrease with higher budget and higher noise. This can be imputed to the reinterpolation step, which may lack robustness in those cases.

The relatively disappointing performances of $EQ90$ (with regard to its complexity) can be explained by two reasons. Firstly, it is, by construction, very dependent on the model uncertainty structure and hence on the parameters estimation [19], which in our setup is most of the time not accurate. Secondly, it is designed to return a solution with small error, which may favor repetitions or clustering instead of exploration, and this benefit is not apparent in an analysis based on the actual response values only.

On average, the AEI criterion seems a good option for our benchmark, since it is several times the best method, and is rarely very bad. As discussed, the plugin of the kriging mean, also used in AEI, is a sensitive option, and the exploration enhancement due to the penalization function (see equation 18) seems also beneficial.

However, another choice can be made based on the user preference, in order to:

- avoid replications: RI
- enhance uncertainty reduction: $EQ90$
- ease implementation: $MQ10$, $PI50$

For all methods, it seems that the results mainly depend on the capacity of kriging to fit the function based on a very small amount of information (small, noisy DOE). When it is the case (Hartman4, Hartman6, Sphere), all the criteria lead to satisfying results, which means here, considering the difficulty of the optimization setup, an approximate identification of the optimum region. Stronger differences in performances between criteria might appear with different setups, for instance using very large budgets and/or very small noise.

6 Conclusion

The objective of this paper was fourfold: first, make a comprehensive review of the kriging-based methods for

the optimization of noisy functions in a unified framework. Second, compare the different methods based on a benchmark with a large variety of test functions, noises amplitude and computational budget. Finally, identify critical and non-critical factors common to all NKO procedures and state on the overall ability of NKO to solve noisy problems. The results presented here are, of course, not universal, and our conclusions might not apply for very different problems or setups.

Out of the criteria (and variants) detailed in this paper, we found that the two most "natural" alternatives for non kriging experts (sequential minimization of the kriging mean and direct plugin of the minimum of the noisy observations in the EGO algorithm), are poor alternatives in all cases, while the relative performances of the other methods depend on the problem. On average, the augmented expected improvement (AEI) was found as a good alternative, although the choice of the criterion might be different based on user preference (to avoid replications, enhance uncertainty reduction, etc.).

We found here that apart from a small number of exceptions, the choice of the covariance function and the size of the initial DOE were not significant factors regarding the optimization performance. The most significant factor seems here the quality of the covariance parameters estimation, which is a well-known challenge in all kriging-based procedures.

Overall, kriging methods were found as efficient alternatives for optimization in very challenging context: very high noise and very limited number of function evaluations in moderately high dimension, which is, in the authors' opinion, where most benefits can be obtained compared to other classical optimization techniques.

References

1. Ankenman, B., Nelson, B.L., Staum, J.: Stochastic kriging for simulation metamodeling. Operations Research **58**(2), 371–382 (2010). DOI 10.1287/opre.1090.0754
2. Bartz-Beielstein, T., Preuß, M.: Considerations of budget allocation for sequential parameter optimization (SPO). In: L. Paquete, et al. (eds.) Proceedings of the Workshop on Empirical Methods for the Analysis of Algorithms (EMAA 2006), pp. 35–40 (2006)
3. Biermann, D., Weinert, K., Wagner, T.: Model-based optimization revisited: Towards real-world processes. In: Z. Michalewicz, R.G. Reynolds (eds.) Proceedings of the 2008 IEEE Congress on Evolutionary Computation (CEC 2008), 1.-6. June, Hong Kong, pp. 2980–2987. IEEE Press, Piscataway, NJ (2008). DOI 10.1109/CEC.2008.4631199
4. Dixon, L., Szegö, G.: Towards global optimisation 2, vol. 2. North Holland (1978)
5. Emmerich, M.: Single- and multi-objective evolutionary design optimization assisted by gaussian random field metamodels. Ph.D. thesis, Universität Dortmund (2005)
6. Fernex, F., Heulers, L., Jacquet, O., Miss, J., Richet, Y.: The MORET 4B Monte Carlo code - New features to treat complex criticality systems. In: M&C International Conference on Mathematics and Computation Supercomputing, Reactor Physics and Nuclear and Biological Application, Avignon, France (2005)
7. Forrester, A., Keane, A., Bressloff, N.: Design and Analysis of "Noisy" Computer Experiments. AIAA journal **44**(10), 2331 (2006)
8. Ginsbourger, D., Picheny, V., Roustant, O., Richet, Y.: A new look at Kriging for the Approximation of Noisy Simulators with Tunable Fidelity. In: 8th ENBIS conference, Athens, Greece (2008)
9. Huang, D., Allen, T., Notz, W., Miller, R.: Sequential kriging optimization using multiple-fidelity evaluations. Structural and Multidisciplinary Optimization **32**, 369–382 (2006)
10. Huang, D., Allen, T., Notz, W., Zeng, N.: Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of Global Optimization **34**(3), 441–466 (2006)
11. Iooss, B., Lhuillier, C., Jeanneau, H.: Numerical simulation of transit-time ultrasonic flowmeters: uncertainties due to flow profile and fluid turbulence. Ultrasonics **40**(9), 1009–1015 (2002)
12. Jin, R., Chen, W., Simpson, T.: Comparative studies of metamodelling techniques under multiple modelling criteria. Structural and Multidisciplinary Optimization **23**, 1–13 (2001)
13. Jones, D.: A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization **21**(4), 345–383 (2001)
14. Jones, D., Schonlau, M., Welch, W.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization **13**(4), 455–492 (1998)
15. Kleijnen, J.: Design and analysis of simulation experiments, vol. 111. Springer Verlag (2007)
16. Krige, D.: A statistical approach to some basinc mine valuation problems on the witwatersrand. Journal of the South African Institute of Mining and Metallurgy **52**, 141 (1952)
17. Li, W., Huyse, L., Padula, S.: Robust airfoil optimization to achieve drag reduction over a range of mach numbers. Structural and Multidisciplinary Optimization **24**(1), 38–50 (2002)
18. Matheron, G.: Le krigeage universel. Cahiers du centre de morphologie mathématique **1** (1969)
19. Picheny, V., Ginsbourger, D., Richet, Y.: Noisy expected improvement and on-line computation time allocation for the optimization of simulators with tunable fidelity. In: 2nd International Conference on Engineering Optimization, September 6-9, 2010, Lisbon, Portugal (2010)
20. Ponweiser, W., Wagner, T., Biermann, D., Vincze, M.: Multiobjective optimization on a limited amount of evaluations using model-assisted $\mathcal{S}$-metric selection. In: G. Rudolph, T. Jansen, S. Lucas, C. Poloni, N. Beume (eds.) Proceedings of the 10th International Conference on Parallel Problem Solving from Nature (PPSN), 13.-17. September, Dortmund, no. 5199 in Lecture Notes in Computer Science, pp. 784–794. Springer, Berlin (2008)
21. Rasmussen, C., Williams, C.: Gaussian processes for machine learning. MIT Press (2006)
22. Roustant, O., Ginsbourger, D., Deville, Y.: The DiceKriging package: kriging-based metamodeling and optimization for computer experiments. Book of abstract of the R User Conference (2009)

23. Sacks, J., Welch, W., Mitchell, T., Wynn, H.: Design and analysis of computer experiments. Statistical science pp. 409–423 (1989)
24. Sakata, S., Ashida, F.: Ns-kriging based microstructural optimization applied to minimizing stochastic variation of homogenized elasticity of fiber reinforced composites. Structural and Multidisciplinary Optimization **38**, 443–453 (2009)
25. Sakata, S., Ashida, F., Zako, M.: Microstructural design of composite materials using fixed-grid modeling and noise-resistant smoothed kriging-based approximate optimization. Structural and Multidisciplinary Optimization **36**, 273–287 (2008)
26. Santner, T., Williams, B., Notz, W.: The design and analysis of computer experiments. Springer (2003)
27. Sasena, M.: Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations. Ph.D. thesis, University of Michigan (2002)
28. Sekhon, J., Mebane, W.: Genetic optimization using derivatives. Political Analysis **7**(1), 187 (1998)
29. Simpson, T., Booker, A., Ghosh, D., Giunta, A., Koch, P., Yang, R.J.: Approximation methods in multidisciplinary analysis and optimization: a panel discussion. Structural and Multidisciplinary Optimization **27**, 302–313 (2004)
30. Vazquez, E., Villemonteix, J., Sidorkiewicz, M., Walter, É.: Global optimization based on noisy evaluations: an empirical study of two statistical approaches. In: Journal of Physics: Conference Series, vol. 135, p. 012100 (2008)