A benchmark of kriging-based infill criteria for noisy optimization

Victor Picheny Tobias Wagner David Ginsbourger

December 2012

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 review of existing kriging-based methods for the optimization of noisy functions is provided. In summary, ten methods for choosing the sequential samples are described using a unified formalism. They are compared on analytical benchmark problems, whereby the usual assumption of homoscedastic Gaussian noise made in the underlying models is meet. Different problem configurations (noise level, maximum number of observations, initial number of observations) and setups (covariance functions, budget, initial sample size) are considered. It is found that the choices of the initial sample size and the covariance function are not critical. The choice of the method, however, can result in significant differences in the performance. In particular, the three most intuitive criteria are found as poor alternatives. Although no criterion is found consistently more efficient than the others, two specialized methods appear more robust on average.

1 Introduction

The use of kriging for modeling and optimizing deterministic computer simulations has a long and successful tradition [30, 20, 37, 21]. In recent years, there has been an increasing interest in the study of "stochastic" simulators, whose outputs can only be observed in the presence of noise. Examples of such simulators can be found in a wide range of applications, including nuclear safety assessment [9], discrete event simulation [1], acoustic wave propagation in turbulent fluids [17], airfoil optimization [23], design of composite materials [32, 31] and experimental measurements in mechanical engineering [4].

The variety of applications has resulted in different approaches for *Noisy Kriging-based Optimization* (NKO) over the last years [10, 15, 39, 25, 26, 38, 35]. Thereby, most NKO algorithms use the same formulation of the kriging model. Consequently, their differences mainly base on different variants of the criterion for selecting the next evaluation point(s) – the so-called infill sampling 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 been only compared to state-of-the-art approaches in their respective fields, but not between them.

The class of optimization problem addressed by most NKO algorithms usually takes the form of a deterministic objective function $y : \mathbf{x} \in D \subset \mathbb{R}^d \to y(\mathbf{x}) \in \mathbb{R}$ with box-constrained decision space, where experiments can only provide noisy observations $\widetilde{y_i} = y(\mathbf{x}^i) + \epsilon_i$ of the true responses $y(\mathbf{x}^i)$ $(1 \leq i \leq n)$. Such a general formulation covers a wide variety of scenarii that can be encountered in real-world applications. A complete coverage of all possible subdomains is thus not possible. For instance, the performances of NKO strategies depends on the nature of the noise (Gaussian or not, homo- or heteroskedastic, autocorrelated, correlated or not with the response), the nature of the objective function (regularity, uni- and multimodality, search space dimensions) or on the cost of observation restricting the experimental budget.

In this paper, we focus on a comprehensive review and benchmark of the different NKO algorithms considering the nature of the objective function and the effect of the experimental budget. This is accomplished based on a set of analytical test functions covering important problem properties, such as uni- and multi-modality, low and moderate search space dimensions, and different noise levels (variances). In order to keep the experiments within a realistic amount, we restrict our investigation to a particular noise model. The ϵ_i's are considered as being random realizations of i.i.d Gaussian noise variables. The observed values might hence differ for different measurements of y at the same $\mathbf{x}$. As kriging models rely on Gaussian processes, this noise model perfectly copes with the model assumptions. The approaches are thus evaluated under optimal conditions. In this context, it has been shown that many practical problems can be reduced to this particular case by applying suitable transformations [41]. Based on these assumptions, a comparison to non kriging-based approaches would be unfair and is thus out of the scope of the present article. A review of alternative approaches is performed in [3].

Based on the benchmark, the relative strengths and drawbacks of the different NKO approaches are disclosed. 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

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. The computational amount required for allowing all these parameters to be considered in order to obtain reliable and (relatively) case-independent conclusions in a noisy context is only possible due to the use of fast-to-evaluate test functions. An artificial perturbation of a more complex deterministic simulator is thus not considered.

Before the design and the results of the benchmark are presented, the kriging model is described. Then, 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 [24] is a functional approximation method originally coming from geosciences [22], and having been popularized in the computer experiments [30] and machine learning [28] communities. In such frameworks, Kriging is often build upon the assumption that the objective function y is one realization of a Gaussian random field Y. Here, we particularly focus on *Ordinary Kriging* (OK), which is described in the following.

2.1 Ordinary kriging

In the OK framework, the random field Y is commonly 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 centered Gaussian field (or *process*) with stationary (translation-invariant) covariance kernel $k : (\mathbf{x}, \mathbf{x}') \in D^2 \to k(\mathbf{x}, \mathbf{x}') = \sigma^2 r(\mathbf{x} - \mathbf{x}'; \psi)$, where r is an admissible correlation function with parameters ψ.

Under such hypotheses [1], the Kriging model can be written in terms of conditional expectation and variance of Y knowing the observations:

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

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

The function m 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, which can be seen as a pointwise quantification of the prediction uncertainty. After conditioning on n observations, the OK mean and variance functions 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 is a weighted sum of the y_i's:

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

with $\boldsymbol{\lambda}^n(\mathbf{x}) = \left(\mathbf{k}_n(\mathbf{x})^T + \frac{(1 - \mathbf{k}_n(\mathbf{x})^T \mathbf{K}_n^{-1} \mathbf{1}_n)}{\mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{1}_n} \mathbf{1}_n^T\right) \mathbf{K}_n^{-1}$. Furthermore, kriging can be characterized in terms of its established properties as *best linear unbiased predictor*.

2.2 Kriging with noisy observations

In the framework of noisy observations, the $\widetilde{y}_i$ can be considered as realizations of random variables $\widetilde{Y}_i := Y(\mathbf{x}^i) + \varepsilon_i$, and Kriging amounts to conditioning Y on the noisy observations $\tilde{y}_i$ $(1 \leq i \leq n)$. As shown earlier in [11], provided that Y and the Gaussian measurement errors ε_i are stochastically independent, the process Y is still Gaussian conditionally on the noisy observations $\tilde{y}_i$ $(1 \leq i \leq n)$. Its conditional mean and variance functions are given by similar OK equations, with the 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 $\widehat{\mu}_n$ fomula, where τ^2 is the variance of the noise variables ε_i.

In computer experiments, the alternative formulation $\widetilde{\mathbf{R}}_n = (\sigma^2 + \tau^2)^{-1} \widetilde{\mathbf{K}}_n$ 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 commonly called *nugget* and $\nu = \frac{\tau^2}{\sigma^2 + \tau^2}$ *scaling factor* [34]. Note that an equivalent formulation, using variograms, was called *ns-kriging* in [31].

In the case of heterogeneous noise variances, i.e. when the $\tau_i^2 := var(\tilde{Y}_i)$ are not all equal, $\tau^2 \mathbf{I}_n$ is replaced by $diag(\left[\tau_1^2 \ldots \tau_n^2\right])$ [11, 42]. In our framework, the observation noise is homoscedastic, but a generalized model is used for the EQI criterion computation (see 3.5). Note that heterogeneous noise variances could be used to handle repetitions and reduce the size of the covariance matrix of the model. For the sake of brevity, we consider here that we have one observation for each input set, but the $\mathbf{x}^i$ are not necessarily all distinct.

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

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. Therewith, the model presented in the paper slightly differs from the so-called *kriging with nugget effect* of the geostatistics literature [24, 6], where τ^2 also appears in the covariance vector $\mathbf{k}_n(\mathbf{x})$ (when $\mathbf{x} = \mathbf{x}_i$), which makes it a discontinuous, interpolating model.

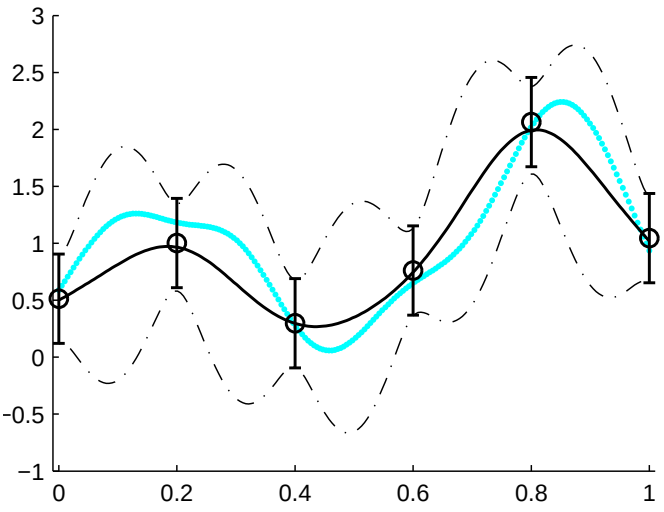


Figure 1: Actual function (bold gray), Kriging mean (bold black) and 90% confidence intervals (mixed line); the circles are the observation values $\widetilde{y}_i$, the bars show the noise amplitude ($\pm 2\tau$).

2.3 Covariance functions

A large variety of covariance kernels are available in the literature (see , e. g., [33] or [28] 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:

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

- the Matérn tensor-product kernel with $\nu = 3/2$:

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

with $\mathbf{x} = [x_1 \dots x_d]$. Both kernels depend on a set of parameters, σ^2 and $\{\theta_1, \dots, \theta_d\}$, which are often referred to respectively as *process variance* and *ranges*.

2.4 Covariance parameter estimation

Covariance parameters are usually estimated based on the observation vector $\widetilde{\mathbf{y}}^n$. To accomplish this, several methods are available, e. g., maximum-likelihood approaches, variogram estimation, 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 of $\widetilde{\mathbf{y}}^n$ seen as a function of the covariance parameters, under a Gaussian assumption on $\widetilde{\mathbf{Y}}^n$:

$$\begin{aligned} L &= (2\pi)^{-\frac{n}{2}} \det\left[\widetilde{\mathbf{K}}_n\right]^{-\frac{1}{2}} \\ &\quad \exp\left(-\frac{1}{2}\left(\widetilde{\mathbf{y}}^n - \widehat{\mu}_n \mathbf{1}_n\right)^T \widetilde{\mathbf{K}}_n^{-1}\left(\widetilde{\mathbf{y}}^n - \widehat{\mu}_n \mathbf{1}_n\right)\right) \end{aligned} \qquad (9)$$

or equivalently by minimizing a negative multiple of the log-likelihood (omitting constants):

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

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" or "profile" log-likelihood*). Unfortunately, the superimposed 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:

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

where the dependency of l on the parameters appears through $\widetilde{\mathbf{K}}_n$ and $\widehat{\mu}_n$. If τ^2 is unknown, it can also be considered as a variable in the likelihood maximization. Here, we set it to the known variance of the superimposed homoscedastic noise according to the idea of benchmarking the approaches under ideal conditions.

3 Infill criteria

In the sequential procedure of most kriging-based optimizers, the design point to be evaluated next is determined based on the optimization of a so-called infill criterion. This infill criterion uses information of the current metamodel in order to assess the utility of evaluating any candidate design on the actual problem. In this section, we present the definitions of and the ideas behind the infill criteria analyzed in our benchmark.

3.1 The classical (noiseless) 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. [20]. Let

$$I_n(\mathbf{x}) := \left(\min(Y(\mathbf{X}^n)) - Y(\mathbf{x})\right)^+ \qquad (12)$$

denote the *improvement* obtained by evaluating Y at $\mathbf{x}$ after the n^{th} iteration, where $(.)^+ := \max(0,.)$, $\mathbf{X}^n = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ and $Y(\mathbf{X}^n) = \left(Y(\mathbf{x}^1), \ldots, Y(\mathbf{x}^n)\right)^T$ (same convention is used later for $\widetilde{Y}(\mathbf{X}^n)$, $m_n(\mathbf{X}^n)$ and $y(\mathbf{X}^n)$.

EI_n is defined as the expectation of I_n conditionally on the observations:

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

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

As shown in [20], the EI is fortunately analytically tractable

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

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

In the EGO algorithm, the next measurement is performed where EI is maximum:

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

By construction, EI_n is always non-negative, strictly increasing with s_n and decreasing with ${m_n}^2$. Furthermore, 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.

3.2 Expected improvement with "plug-in" (PI)

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 representative of y_{min}. This leads to the so-called *Expected Improvement with plugin*, denoted here by $EI_{T,n}$:

$$EI_{T,n}(\mathbf{x}) = \mathbb{E}[(T - Y(\mathbf{x}))^+ | \widetilde{Y}(\mathbf{X}^n) = \tilde{\mathbf{y}}^n] \quad (16)$$

The choice of T is an important issue, since too high or too low values have a significant influence on the shape of $EI_{T,n}$ and thus change its behavior relatively to EI with known y_{min} [19]. A first "naive" approach consists in choosing $T = \min(\tilde{\mathbf{y}}^n)$, but this plugin lacks robustness since it suffices to have one noisy observation with a low value to severely underestimate y_{min} for the rest of the optimization. Following the approach mentioned in [39], $T = \min(m_n(\mathbf{X}^n))$ 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.

In [25], an almost similar EI with plugin was proposed (in a bayesian kriging framework, which we do not consider here): $T = \min_{s_n(\mathbf{x}^i) \leq e}\left(m_n(\mathbf{x}^i)\right)$, for some parameter e, in order to restrict the choice to the observations with relatively high accuracy.

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:

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

However, one drawback of $EI_{T,n}$ for noisy optimization is that it does not take into account the noise of the future observation: the improvement is defined and its expectation 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_n^2 is small compared to the noise level τ^2.

3.3 Augmented Expected Improvement (AEI)

The AEI criterion was proposed by Huang et al. ([15] for the noisy framework and [14] for multi-fidelity). The idea of replacing the unknown y_{min} by the value of the kriging mean at some point is also used. But this time, instead of considering $T = \min(m_n(\mathbf{X}^n))$, 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 plugin less sensitive to noise. In other words, T is the kriging mean value at the design point with lower β-quantile, where $\Phi^{-1}(\beta) = \alpha$. The value $\alpha = 1$ (corresponding with $\beta = 0.84$) is recommended by the authors.

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

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

AEI 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 (RI)

The *reinterpolation* method was proposed by Forrester et al. [10]. Instead of modifying the EI criterion for the noisy

[2] We consider minimization problems in this paper.

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 parameters of the regressing model, as well as the design of experiments (DOE), 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 since $EI\left(\mathbf{x}^i\right) = 0$, $1 \leq i \leq n$.

3.5 Expected Quantile Improvement (EQI)

The main idea behind the EQI criterion, as detailed in [26], is that in a noisy situation, the model predictor may be closer to the actual value than the original data; hence, an improvement should refer to the effect of a new observation on the model. Taking the kriging quantile $q_n(\mathbf{x}) = m_n(\mathbf{x}) + \Phi^{-1}(\beta)s_n(\mathbf{x})$ (with $\beta \in [0.5, 1[$) as a measure of reference, an improvement between steps n and $n+1$ is defined, similarly to the noiseless case, as:

$$I_n(\mathbf{x}) := \left(\min_{1 \leq i \leq n} (Q_n(\mathbf{x}^i)) - Q_{n+1}(\mathbf{x}) \right)^+ \tag{19}$$

where Q_{n+1} is the quantile of the kriging updated with a new measurement at $\mathbf{x}^{n+1} = \mathbf{x}$. EQI is defined as $\mathbb{E}[I_n(\mathbf{x})|\widetilde{Y}(\mathbf{X}^n) = \tilde{\mathbf{y}}^n]$, the expectation being taken conditionally on the n current measurements and on the fact that a new measurement is made at $\mathbf{x}^{n+1}$, the new measurement $\widetilde{Y}(\mathbf{x}^{n+1})$ being (at step n) a Gaussian random variable with mean $m_n(\mathbf{x}^{n+1})$ and variance $s_n^2(\mathbf{x}^{n+1}) + \tau_{n+1}^2$. It has been shown that the conditional distribution of $Q_{n+1}(\mathbf{x})$ is Gaussian and analytically derivable, which leads to the following formula for the EQI:

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

where $q_{min} := \min_{1 \leq i \leq n}(q_n(\mathbf{x}^i))$ 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, m_Q and s_Q take the simple following form:

$$m_Q(\mathbf{x}) = m_n(\mathbf{x}) + \Phi^{-1}(\beta)\sqrt{\frac{\tau_{new}^2 s_n^2(\mathbf{x})}{\tau_{new}^2 + s_n^2(\mathbf{x})}} \tag{21}$$

$$s_Q^2(\mathbf{x}) = \frac{\left[s_n^2(\mathbf{x})\right]^2}{\tau_{new}^2 + s_n^2(\mathbf{x})} \tag{22}$$

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 along the iterations.

3.6 Approximate knowledge gradient (AKG)

The knowledge-gradient policy that has been proposed in [35] aims at measuring the *global* effect of a new measurement on the kriging mean. The *knowledge improvement* is defined as the difference between the minima over D of the functions m_n and m_{n+1}. Since it requires two global searches over D, no closed-form equation for continuous optimization exists, but an efficient approximation has been proposed, termed approximated knowledge gradient (AKG). The knowledge improvement is then defined as:

$$I_n(\mathbf{x}) = \min\left[M_n\left(\mathbf{X}^{n+1}\right)\right] - \min\left[M_{n+1}\left(\mathbf{X}^{n+1}\right)\right] \tag{23}$$

where $\mathbf{X}^{n+1} = \{\mathbf{X}^n, \mathbf{x}\}$, and M_{n+1} denotes the mean of the kriging updated with a measurement at $\mathbf{x}^{n+1} = \mathbf{x}$. Note that $M_{n+1} = Q_{n+1}$ for $\beta = 0.5$, so conditionally for the n first observations M_{n+1} is a Gaussian process with known distribution.

In other words, an improvement is obtained if the minimum of the new kriging mean is smaller that the minimum of the old kriging mean, both minima being taken over the $n+1$ sample points. In contrast, EQI considers the minimum of the old kriging quantile over the n past sample points, and the new kriging quantile at the new sample point only.

AKG is defined as $\mathbb{E}[I_n(\mathbf{x})|\widetilde{Y}(\mathbf{X}^n) = \tilde{\mathbf{y}}^n; \mathbf{x}^{n+1} = \mathbf{x}]$. The difficulty here lies in the computation of the conditional expectation of $\min\left[M_{n+1}\left(\mathbf{X}^{n+1}\right)\right]$, since it is the minimum of a Gaussian vector. However, the problem can be reformulated as

$$\min\left[M_{n+1}\left(\mathbf{X}^{n+1}\right)\right] = \min_{i \in \{1,\ldots,n+1\}}\left[m_n(\mathbf{x}^i) + s_M(\mathbf{x}^i)Z\right] \tag{24}$$

where Z is standard normal (scalar), and:

$$s_M(\mathbf{x}^i) = \frac{c_n(\mathbf{x}^i, \mathbf{x})}{\sqrt{s_n^2(\mathbf{x}) + \tau^2}} \tag{25}$$

with

$$c_n(\mathbf{x},\mathbf{y}) = k(\mathbf{x},\mathbf{y}) - \mathbf{k}_n(\mathbf{y})^T\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}) + \frac{\left(1-\mathbf{k}_n(\mathbf{y})\mathbf{K}_n^{-1}\mathbf{1}_n\right)\left(1-\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})\right)}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n} \quad (26)$$

Then, the expectation of the maximum takes the following form:

$$\mathbb{E}\left[\min M_{n+1}\left(\mathbf{X}^{n+1}\right)|\widetilde{Y}(\mathbf{X}^n)=\tilde{\mathbf{y}}^n;\mathbf{x}^{n+1}=\mathbf{x}\right] = \sum_{i=1}^{\tilde{n}}\tilde{a}_i\left(\Phi\left(\tilde{c}_{i+1}\right)-\Phi\left(\tilde{c}_i\right)\right)+\tilde{b}_i\left(\phi\left(\tilde{c}_i\right)-\phi\left(\tilde{c}_{i+1}\right)\right) \quad (27)$$

where $\tilde{a}$, $\tilde{b}$ and $\tilde{c}$ are determined by running a small algorithm (see Table 1 in [35]).

3.7 Minimal quantile criteria (MQ)

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:

$$\mathbf{x}^{n+1} = \arg\min_{\mathbf{x}\in D} m_n(\mathbf{x}) + \Phi(\beta)^{-1}\times s_n(\mathbf{x}) \quad (28)$$

This criterion can be found in [5] or [38]. In [38], $\Phi(\beta)^{-1}$ is an increasing function of the number of iterations in order to ensure asymptotic convergence.

Although recognized as less efficient compared to the *EI* in the case of deterministic experiments [19], this method seems worth studying in this benchmark because it does not need any modification to handle noise. It also has shown successful applications in kriging-based multi-objective optimization [8, 27].

3.8 Integral criteria

Alternatively to pointwise-based criteria, two strategies have been proposed that rely on *global* measures of improvement: the Informational Approach to Global Optimization (IAGO) [40] and the Integrated Expected Conditional Improvement (IECI) [12]. The IAGO strategy maximizes the gain of information, i.e. selects for the next evaluation the point that minimizes the expected conditional entropy of the minimizer; the IECI evaluates by how much a candidate measurement at a given point would affect the expected improvement over the design space. Both strategies can handle noise naturally. The major drawback of such criteria is that they cannot be computed in closed form and rely on numerical integration (contrarily to the other criteria presented here, for which the value, and also the derivatives, can be calculated in an exact way, whatever the dimension). Comparing pointwise and integral criteria brings up additional questions about the choice of the integration method in moderate/high dimension (which can become very expensive rapidly). For those reasons, we did not include these two criteria into the current benchmark.

4 Design of the benchmark

4.1 Analytical test functions

As test problems, we employed six widely used analytical benchmark problems [7]. Their definitions are given in Table 1. The original functions have been rescaled to map their search space to $D = [0,1]^d$, their mean to zero, and their variance to one (for a random design with uniform distribution over D). 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 and dimensions. *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. *Branin-Hoo, Goldstein-Price, Hartman4* and *Hartman6* are multimodal functions with a moderate number of optima. In addition, the smoothness of the function differs from a constant curvature (Sphere6) to a strong bend response surface with a steep optimum region (Goldstein-Price).

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 noise level and the allowed budget of evaluations (a similar classification can be found in [18]). For the tuning factors, we selected the ones which have been changed within different studies [2, 4], i. e., the proportion of observations for the initial DOE and the choice of the covariance kernel.

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 of all these parameters (including the infill criterion), results 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 $n_{fct} \times n_{noises} \times n_{criteria} \times n_{cov} \times n_{budgets} \times n_{DOEsizes} \times n_{runs}$=63,360.

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)	$y(\mathbf{x}) = \frac{1}{2.427}\left[\log\left[\left(1 + (\bar{x}_1 + \bar{x}_2 + 1)^2(19 - 14\bar{x}_1 + 3\bar{x}_1^2 - 14\bar{x}_2 + 6\bar{x}_1\bar{x}_2 + 3\bar{x}_2^2)\right)\left(30 + (2\bar{x}_1 - 3\bar{x}_2)^2\left(18 - 32\bar{x}_1 + 12\bar{x}_1^2 + 48\bar{x}_2 - 36\bar{x}_1\bar{x}_2 + 27\bar{x}_2^2\right)\right)\right] - 8.693\right]$ with: $\bar{\mathbf{x}} = 4 \times \mathbf{x} - 2$
Rosenbrock4 (4D)	$y(\mathbf{x}) = \frac{1}{3.755\times10^5}\left[\sum_{j=1}^{3}\left(100(\bar{x}_{j+1} - \bar{x}_j^2)^2 + (1 - \bar{x}_i)^2\right) - 3.827 \times 10^5\right]$ with: $\bar{\mathbf{x}} = 15 \times \mathbf{x} - 5$
Hartman4 (4D)	$y(\mathbf{x}) = \frac{1}{0.839}\left[1.1 - \sum_{i=1}^{4} C_i \exp\left(-\sum_{j=1}^{4} a_{ji}\left(x_j - p_{ji}\right)^2\right)\right]$ 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)	$y(\mathbf{x}) = \frac{-1}{1.94}\left[2.58 + \sum_{i=1}^{4} C_i \exp\left(-\sum_{j=1}^{6} a_{ji}\left(x_j - p_{ji}\right)^2\right)\right]$
Sphere6 (6D)	$y(\mathbf{x}) = \frac{1}{899}\left[\sum_{j=1}^{6} x_j^2 \times 2^j - 1745\right]$

Table 2: Summary of the benchmark factors and levels.

Factor	**Values**
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

Table 3: Summary of the infill criteria.

Criterion	**Parameter**	Abbreviation
Random search	-	RS
Reinterpolation	-	RI
AEI	$\alpha = 1$	AEI
EQI	$\beta = 0.5$	EQ50
	$\beta = 0.9$	EQ90
EI with plugin	$T = \min\left(\tilde{y}^i\right)$	PIy
	$T = \min\left(m_n(\mathbf{X}^n)\right)$	PI50
	$T = \min\left(q_n(\mathbf{X}^n)\right)$ with $\beta = 0.9$	PI90
Quantile minimization	$\beta = 0.5$	MQ50
	$\beta = 0.1$	MQ10
AKG	-	AKG

The choice of the design type of the initial DOE is also probably a significant factor; however, here we fix it to be a latin hypercube sampling (LHS) 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 (MQ) depend on the quantile level β. For those methods, two levels ($\beta = 0.5$ and $\beta = 0.9$) are tested. Since it is almost similar to using $\beta = 0.5$, the plugin of [25] is not considered here. Using a varying β for quantile minimization as in [38] is critical mostly in asymptotic conditions, while in our benchmark we consider only small numbers of observations. Besides, no rule-of-thumb is provided for tuning β. For those reasons, we limit our present study to the fixed β case. 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 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.

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* [29]. 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. 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.

Since *Rosenbrock4* and *Sphere6* have a very low activity, the covariance bounds for the range parameters are chosen as $[0.5, 5]$, which allows to have a very smooth kriging model. For *Branin-Hoo, Goldstein-Price, Hartman4* and *Hartman6*, the range bounds are set to $[0.1, 1]$, which allows to model high activity responses.

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 may fail due to numerical instability in the inversion of the covariance matrix. When this occurs, the model is updated based the old covariance parameters. In particular, the reinterpolation technique is 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 feasible 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 illustrate 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.

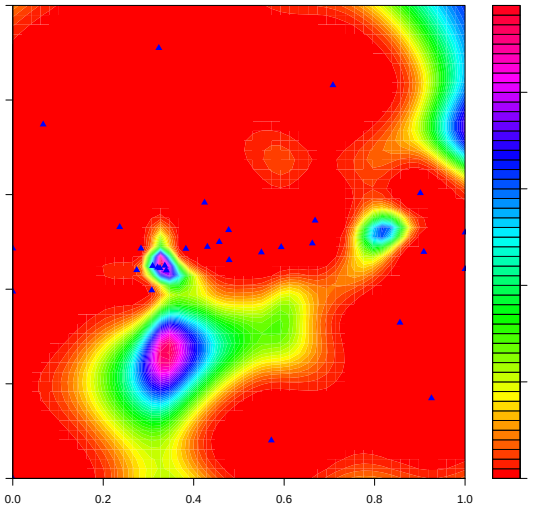


Figure 2: Contour lines of the AEI during a typical optimization run (Goldsteinprice function, 29 points (triangles), noise level 20%).

Here, we chose to optimize the infill criteria using the *genoud* algorithm (GENetic Optimization Using Derivatives, [36]), 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* [29]. 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 values presented in Table 4.3.2. 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,

Table 4: Parameterization of the genoud algorithm for infill criteria optimization..

Parameter	**Name in the control variable of genoud**	**value**
Population size	`pop.size`	6×2^d
Evolutionary generations	`max.generations`	20
Maximum evaluations within a gradient descent	`BFGSmaxit`	6×2^d
Target tolerance of the optimization	`solution.tolerance`	0
Maximum generations without improvement	`wait.generations`	2
Evolutionary generations before first gradient descent	`BFGSburnin`	0

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.

5 Results

5.1 Analysis of average performance using ANOVA

5.1.1 Procedure

We start by analyzing the results quantitatively using the true objective value $y(\mathbf{x}^*)$ of the design, which is identified as the current best by the corresponding infill criterion for each run. In more detail, for PIy, $\mathbf{x}^*$ is the design point with the best noisy observation, for AKG, RI, PI50, EQ50, MQ50 and RS, $\mathbf{x}^*$ is the design point with the best kriging mean, and for AEI, PI90 and EQ90, $\mathbf{x}^*$ is the one with the best kriging quantile.

In order to analyze the very large amount of data over different configurations, we proceed in two steps. First, we use ANOVA to analyze the effects of the factors, provide a first comparison of the methods and remove the least efficient ones. Then, detailed results are provided in the form of boxplots.

We start our analysis by a screening phase, and consider the significance of the effects of the different algorithm parameters. The full factorial structure of the benchmark allows the effects of each parameter to be estimated in an unbiased way, i. e., the change of performance obtained by changing a parameter can be assessed over all combinations of the remaining parameters. To do so, we postulate, for each test function, a linear model with second order interactions for the performance response, in the fashion of [16]. Each factor (noise level, budget, initial DOE size, covariance, LHS instance, criterion) is treated as categorical, and the model is generated using deviation coding (see [13], chap. 5). For the LHS, only the main effect is considered in order to integrate out its effect on the performance, e. g., the inclusion of a point close to the true optimum. It is thus is used as blocking variable; no interactions with the other factors are considered. Note that such analysis considers average performances only, and, in particular, it does not allow us to assess the robustness of a given setup, which is crucial in optimization. Hence, here we aim at identifying general trends, the other issues being treated in Section 5.2. The performance response is chosen as the logarithm of the difference between the function value at the best point and the actual minimum of the function (log of optimality gap):

$$y = \log(y(\mathbf{x}^*) - y^*). \tag{29}$$

Table 5: Goodness-of-fit statistics of the linear models.

Test	R^2	R^2_{adj}
Branin	0.48	0.47
Goldstein-Price	0.44	0.43
Hartman4	0.62	0.61
Rosenbrock	0.22	0.21
Hartman6	0.66	0.65
Sphere	0.70	0.70

The logarithmic transformation is used to provide approximately normally distributed residuals to the linear model of the ANOVA. The R^2 and R^2_{adj} values of each model are reported in Table 5. The relatively low values indicate that a large part of the variance of the performance cannot be explained by the model, and is due to the observation noise realizations during the optimization.

The main effects are given on Table 6. The complete ANOVA tables are reported in Appendix B (Tables 7, 8, 9, 10, 11 and 12).

In the following, we organize our observations by the factors tunable by the user: the size of the initial design, the covariance kernel and the infill criterion.

5.1.2 Influence of the initial design size

The initial design size has a significant main effect (Appendix 9) on *Branin*, *Hartman4*, *Rosenbrock* and *Sphere*, for which the large doe is slightly preferred (except for *Rosenbrock* for which the fit of the ANOVA is worse). However, we see that, on average (Table 6), there is very little difference between the small and large initial DOE

Table 6: Main effect of the factors for each test function. Best results among methods are in bold fonts.

	Branin	Goldstein-Price	Hartman4	Rosenbrock	Hartman6	Sphere
Constant	-1.69	-0.64	-1.00	-1.89	-1.13	-1.24
budget=20	0.09	0.13	0.09	0.03	0.05	0.05
budget=40	-0.09	-0.13	-0.09	-0.03	-0.05	-0.05
noise=5	-0.61	-0.42	-0.47	-0.22	-0.35	-0.46
noise=20	0.10	0.01	0.06	0.05	0.04	0.05
noise=50	0.52	0.41	0.41	0.17	0.31	0.41
init=4	0.01	0.01	0.01	-0.01	0.00	0.02
init=10	-0.01	-0.01	-0.01	0.01	0.00	-0.02
cov=G	-0.02	0.08	-0.01	0.00	-0.02	0.00
cov=M32	0.02	-0.08	0.01	0.00	0.02	0.00
method=AEI	-0.10	**-0.40**	-0.38	-0.04	-0.14	-0.22
method=AKG	**-0.12**	-0.29	**-0.48**	-0.06	**-0.19**	**-0.24**
method=EQ50	-0.07	-0.06	-0.12	-0.05	-0.11	-0.13
method=EQ90	-0.04	0.14	0.06	-0.10	-0.08	-0.05
method=MQ10	-0.08	-0.11	-0.11	-0.08	-0.14	-0.09
method=MQ50	0.15	0.42	0.54	**-0.12**	0.23	0.26
method=PI50	-0.05	-0.03	-0.03	-0.04	-0.09	-0.04
method=PI90	0.02	0.24	0.25	-0.10	0.00	0.08
method=PIy	0.11	-0.32	-0.11	0.27	0.10	0.07
method=RI	-0.07	-0.15	-0.32	0.09	-0.11	-0.22
method=RS	0.26	0.56	0.72	0.23	0.53	0.59

size. Considering the interactions, particularly the ones with the noise level and criterion are significant. Whereas the effect of the noise level on on the preferred design is different depending on the test functions, one can remark that a smaller design size is slightly better with AKG and AEI, and larger designs are better with MQ50, the other criteria being indifferent. Hence, we can conclude that, for our benchmark, the DOE size has little influence on the optimization results, or its influence is limited to a few particular configurations.

5.1.3 Influence of the covariance kernel

The covariance has a significant main effect (Appendix 9) only on the multimodal functions: *Branin*, *Goldstein-Price*, *Hartman4*, *Hartman6*. Thereby, no clear trend can be observed. Overall the values (Table 6) are small compared to the ones associated with the criteria.

Interactions with noise are significant on *Branin*, *Goldstein-Price*, *Hartman4* and *Rosenbrock*. Interactions with criterion are significant on all test functions, but *Branin*. Hence, these interactions are more closely inspected in the following.

5.1.4 Influence of the infill criterion and interaction with covariance

A first look at Table 6 shows that two methods seem to provide good performances on average: AEI and AKG, and two seem less efficient: MQ50 and PI90. MQ50 performs consistently poorly (except on *Rosenbrock*); hence, for this family of criteria, using a low quantile such as MQ10 seems a lot better alternative. Similarly, it can be noted that on almost all the configurations (except on *Rosenbrock*), PI90 is less efficient than PI50. We can then conclude that the plugin of the kriging mean outperforms the plugin of a high kriging quantile. At this point, the picture is not clear between PI50 and PIy.

Interactions of criteria with noise level, budget and covariance are significant on almost all test functions (Appendix 9), meaning that at least some criteria react differently to a change of these factors, and analyzing the main effects does not allow to rank the methods with precision. In order to measure the effects of interactions, we draw the model predictors and confidence intervals of the linear evaluation models (ANOVA) (Figure 3) for all the combinations of budgets, noise levels, covariances and criteria (except MQ50 and PI90, which are already identified as poor solutions). The effects and interactions of the initial design are averaged out for visual clarity based on the minor effect of this factor.

First, we analyze the effect of the covariance. Overall, we see that, the choice of the kernel is mostly negligible, with several exceptions:

- On *Branin* (Figure 3 A), for most methods the Gaussian covariance seems significantly better with 5% noise.
- On *Goldstein-Price* (Figure 3 B), the Matern covariance (circles) seems consistently better than the Gaussian one (triangles). This effect is particularly strong for the successful criteria (AEI, AKG, PIy, RI) and is getting weaker with 20% and 50% noise.
- On *Hartman4* (Figure 3 C), the effect of the covariance is important only for AKG and RI (Matern being

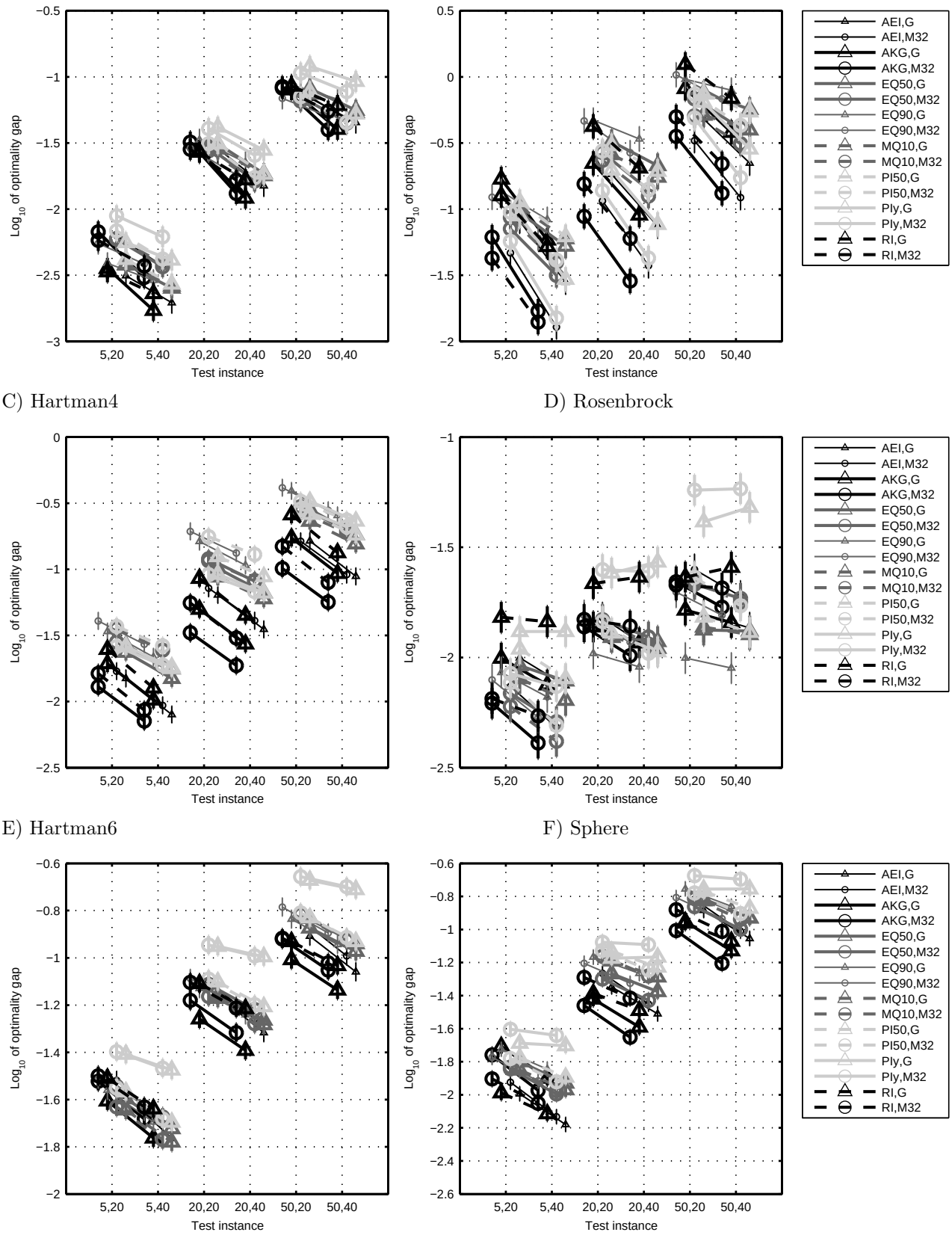


Figure 3: Average performances predicted by the linear model on the six test functions. On the x-axis, the left number represented the noise level and the right number the budget level. The vertical bars show the 95% confidence interval on the predictors. Bars are staggered to improve readability. Increasing the budget or decreasing the noise improve the performance of all methods, which is expected.

slightly better) and PI50 and MQ10, where Gauss is superior.

- On *Hartman6* (Figure 3 E), the Matern covariance is slightly better for AKG with 20% and 50% noise.
- On *Sphere* (Figure 3 F), the Gauss covariance is slightly better for AKG with 20% and 50% noise.

Ranking between the methods is configuration dependent, and no method clearly appears as best here. However, we can observe that:

- AKG, AEI, RI and PIy form a group of more efficient methods on *Goldstein-Price*, especially with the Matern covariance.
- AKG is best on *Hartman4*, regardless of the budget or noise level, and on *Hartman6* and *Sphere* for the 20% and 50% noise levels.
- EQ90 is best on *Rosenbrock* for the 20% and 50% noise levels.

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

- EQ90 and PI50 work poorly on *Goldstein-Price* with high noise and on *Hartman4*.
- With Gaussian covariance, RI works poorly on *Rosenbrock*.
- PIy is the worst method on *Branin, Rosenbrock, Hartman6* and *Sphere*, regardless of the configuration.

Interaction between method and budget is limited to a few configurations. Indeed, the slope between budgets 20 and 40 is almost independent of the noise level and the method, except for the worst and best methods. The effect is particularly strong on *Rosenbrock*, where the slope even becomes positive for PIy and RI, implying that a higher budget results in a worse solution. Inversely, on *Goldstein-Price*, it appears clearly that the best methods tend to make a better use of the additional budget.

5.2 Detailed analysis of infill criteria using boxplots

In this section, we propose a detailed comparison of the methods. Based on the previous observations, we limit now our analysis to the following criteria: AKG, RI, AEI, EQ50, EQ90, PI50 and MQ10. The performance measure is the absolute difference between the function value at the best point ($y(\mathbf{x}^*)$ and the actual minimum of the function (y^*) rescaled by the function standard deviation (σ_y, here always equal to one):

$$D = \frac{|y(\mathbf{x}^*) - y^*|}{\sigma_y} \tag{30}$$

We represent the results in the form of boxplots for each method, test function and noise level. To limit the amount of figures, we show the results for the high budget only, since from Figure 3 it can be seen that the contrasts are slightly higher. We show the results with the Matern covariance and the small initial DOE, which have been found to have little influence of the results, but which provide superior numerical stability and have a small beneficial interaction with the appropriate methods.

On *Branin* (A), with 5% noise, all the criteria accurately identify one of the minima, which is expected since this function is very easy to optimize. With 20% noise, all the median performances are very good, although some outliers can be observed. With 50% noise, all the criteria fail at identifying one of the minima with precision, although with such a high noise the performances can be considered as quite satisfactory. While AKG and AEI were identified as slightly better than the other criteria by the linear model (Table 6) on this function, no differences are measurable by the boxplots in the high noise case. In particular, the performance of AEI improves with increasing noise.

On *Goldstein-Price* (B), with 5% noise, we see that all the median performances but $MQ10$ and $PI50$'s are very good, but heavy tails can also be observed for EQ50 and EQ90. Indeed the basin of the global minimum is relatively small and was not found for several runs. With 20% noise, the difference is clear between AKG, RI and AEI, which almost always capture the global minimum, and the other criteria, which capture it approximately 25% of the time. With 50% noise, results show a high variability, although the criteria identified as best by the linear model (Table 6 and Figure 3 B) perform globally better than the others (AKG and AEI). For this function, the more exploratory criteria are clearly superior, which is expected since the function is multimodal and the optimal region is narrow. One may notice that this effect is much stronger than on the main effect table, as only the results for the Matern covariance are considered.

On *Hartman4* (C), again with 5% noise all the median results are very close to the actual optimum, but only AKG and RI do not have outliers. MQ10 and PI50 perform worst in terms of upper quartile. With 20% and 50% noise, AKG, AEI and RI are globally better than the other methods, and EQ90 and PI50 are slightly worse, which confirms the main effect values.

On *Rosenbrock* (D), differences are difficult to distinguish, as expected from the main effect table. For all methods, the optimum region is identified for more than 75% of the runs for all noise levels. Note that the better performance of EQI detected on Figure 3 D) does not appear, as the results with the Gaussian covariance are not used here.

On *Hartman6* (E), only PIy appears as a poor alternative. The medians are relatively close to the optimum value for all noise levels. With 50% noise, several runs failed at providing a good solution. The global better performance of AKG, indicated by the linear model, is visible in this graph, but also AEI and RI provide satisfactory

results.

On *Sphere* (F), with 5% noise only PI50 and MQ10 are slightly worse than the other methods. With 20% noise, EQ90, PI50 and MQ10 show larger variation. With 50% noise, AKG and RI are slightly better than the other methods with respect to robustness as measured by their upper quartile.

6 Discussion

The first conclusion of this benchmark analysis is the limited influence of the initial design 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 out each other regarding the optimization efficiency. For the user, this choice is thus not a critical one, regardless of the budget, noise, function modality or space dimension.

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 non-critical, except for one specific function (*Goldstein-Price*). Significant interactions appear only with criteria, but tend to be configuration-dependent (e.g., the better performance of EQI and Gauss on *Rosenbrock*. The strongest influence is detected with RI, as the reinterpolation step tends to lack robustness with Gauss, particularly on the smooth functions. In general, Matern may be preferred, as it may provide a better numerical stability for smooth functions and allows a better detection of narrow optimal regions.

The third parameter to be set up by the user is the infill criterion. We found that no criterion outperforms the others on all configurations. However, 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 [19], the question was left open in the noisy case. It is found that the $MQ50$ performances are also poor in presence of noise so this solution is not competitive with other criteria.

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 target that is very easy to reach, which makes 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 on *Goldstein-Price*, for which the minimum is indeed in a small valley, but for most of the configurations it was found inefficient. Note that similar results may be obtained with a low quantile plugin, $PI10$ for instance. Overall the criterion $PI50$ (with an unbiased plugin) appears as a better alternative. This becomes particularly obvious on Rosenbrock where the predicted performance of PIy deteriorates with larger budget, which can be caused by an increased probability of observing an extreme positively biased outlier within more iterations.

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 (at least for the smoothed model), 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 subsequent reinterpolation step, which may lack robustness in those cases.

The relatively disappointing performances of $EQ90$ (with regard to its complexity) can be explained by the fact that 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. An illustration of this characteristic is proposed in Appendix A.

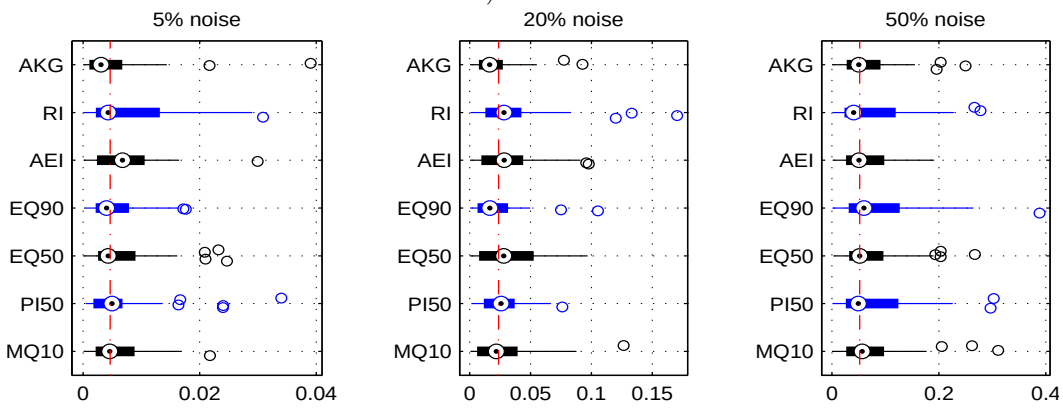
The criteria $MQ10$, $EQ50$, AEI and AKG proved in our context to be competitive; the differences between them depending on the configurations, although no particular pattern regarding the problem parameters (noise, budget, dimension or modality) could be identified by the linear model or the boxplots. The most contrasted results were obtained on *Goldstein-Price* with low and moderate noise, in favor of the most exploratory criteria, as the optimal region is narrow and difficult to detect.

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 sensible option, and the exploration enhancement due to the penalization function (see equation 18) seems also beneficial.

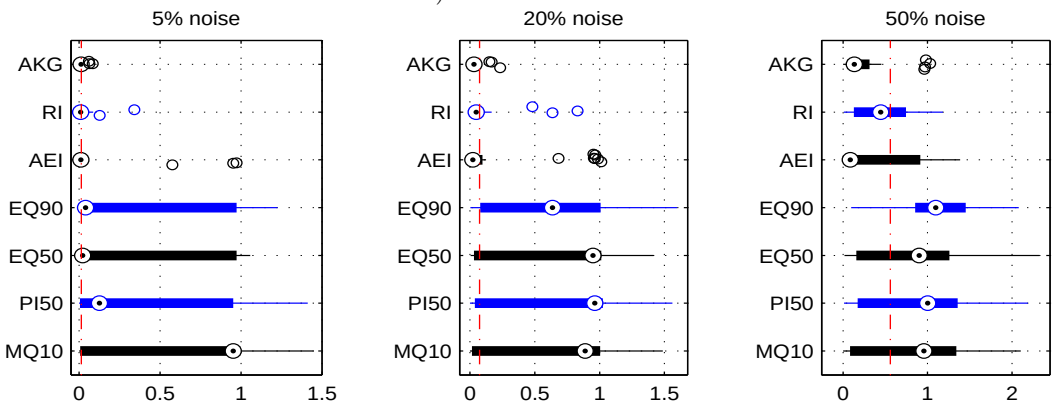
The AKG criterion provided the best performances on several configurations, and is also a sensible choice on average. Two relatively poor performances with small noise (on *Goldstein-Price* and *Sphere*) might indicate that it is more efficient with large noise. Indeed, contrarily to $PI50$, RI, $EQ90$ or AEI, AKG does not reduce to the classical EI in absence of noise.

Overall, one important practical result of this benchmark is that on many configurations, seven of the different methods provide relatively similar results. Hence, if AEI and AKG are identified as slightly better, another choice can be made based on the user preference, in order to

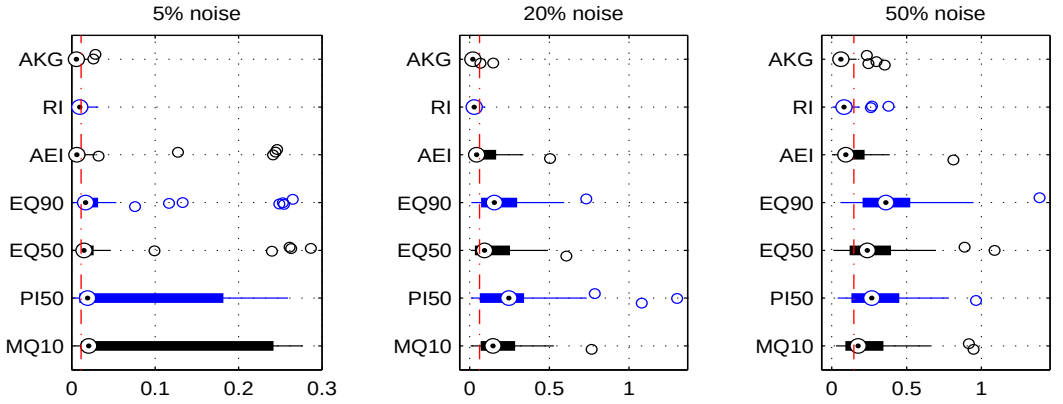
A) Branin:



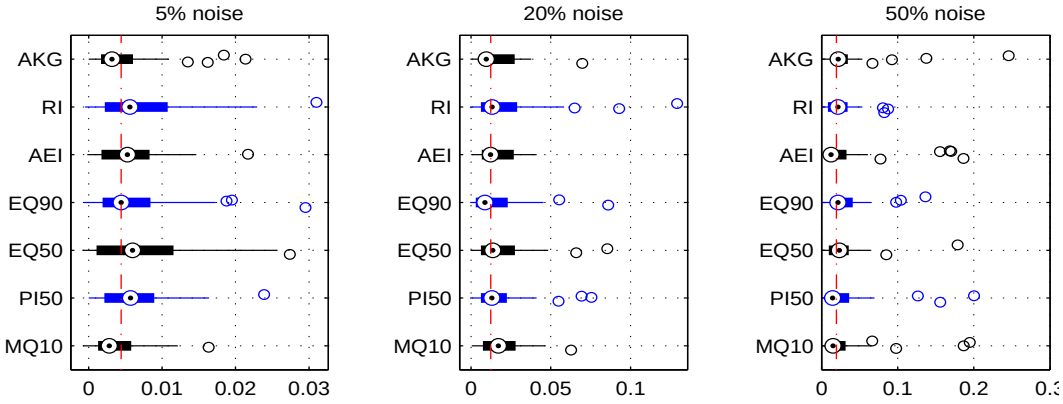
B) Goldstein-Price:



C) Hartman4:



D) Rosenbrock:



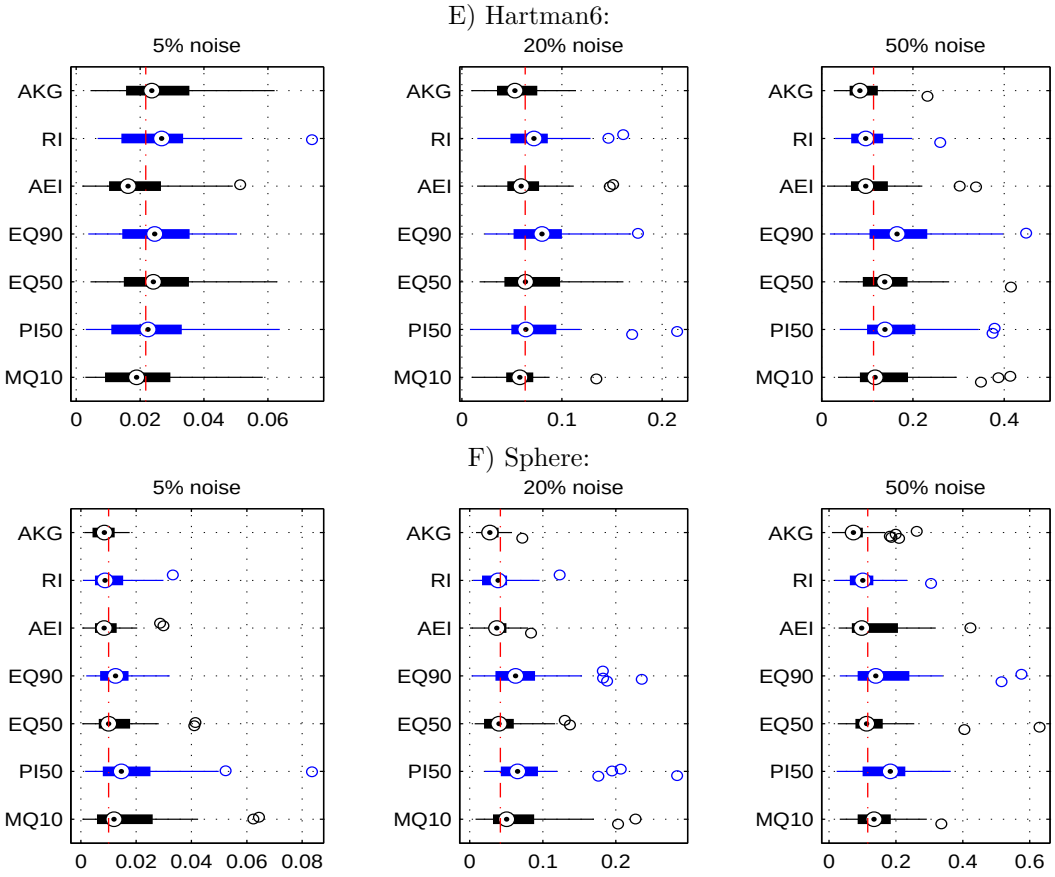


Figure 4: Boxplots of performances (differences between the function value at best point and the actual minimum) for all criteria, test functions and noise levels. The vertical bar shows the overall median performance.

avoid replications (RI) or enhance uncertainty reduction ($EQ90$).

For all methods, it seems that the results mainly depend on the capability 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.

It should be noted that the discussion proposed here is specific to our framework (Gaussian, independent noise with constant known variance), and might vary in a different context. In particular, in the case of known heteroscedastic noise, it is reasonable to conjecture that criteria that account for the noise amplitude (EQI, AKG) will have better performances than the other ones. Inversely, when little information is available for the noise, some criteria might prove to be more robust (RI, $PI50$, $MQ10$).

7 Conclusion

In this research, a comprehensive review of kriging-based methods for the optimization of noisy functions is proposed in a unified framework. The different methods are compared in the case of independent, Gaussian, homoscedastic noise, based on a benchmark of analytical test functions with a large variety of setups that covers a wide range of potential applications. Variations on factors common to all NKO procedures, such as the size of the initial set of experiments or the choice of the covariance kernel, are included in the analysis in order to assess their influence as well as provide a comparison between methods independent of critical arbitrary choices. An extensive experimental design and statistical tools are used to provide a robust and unbiased performance analysis.

First, we found that, apart from a small number of exceptions, the size of the initial DOE is not critical, which means that the effect of having less sequential steps (with larger initial DOE) is counterbalanced by the benefit of a better initial exploration and surrogate model. Hence, the effect of this factor can be neglected without introducing bias in future work. The covariance function is also not a very important factor, which implies that, in the noisy context, the exploration / exploitation trade-off sought for global optimization is achieved regardless of this choice. The numerical stability of the approaches, however, can be often improved by avoiding the Gaussian kernel. In our benchmark, this particularly holds for the RI criterion. If some prior knowledge about the shape of the response surface (smooth or rugged, uni- or multimodal) exists, the kernel should be chosen accordingly, as shown on the *Goldstein-Price* function in our benchmark.

Out of the criteria (and variants) detailed in this paper, we found that several are either poor alternatives ($PI90$, $MQ50$) or lack robustness (PIy). The other criteria have relatively similar performances, although, on average, the augmented expected improvement (AEI) and the approximate knowledge gradient (AKG) were found as best alternatives. In our framework, the practical or statistical considerations that motivated the multiplicity of the criteria seem dominated by the kriging modeling error due to the very small amount of information available. Hence, the choice of the criterion may be based on user preference (to avoid replications, enhance uncertainty reduction, improve implementation robustness, etc.) without a critical deterioration of the performance.

Future research may address the evaluation of the NKO algorithms performances for heteroscedastic noise scenarii, such as variance depending on input parameters (independently of the response), or noise linearly correlated with the response, which are likely to be found in real word problems.

Acknowledgements

The contributions of Tobias Wagner to this paper are based on investigations of the project D5 of the Collaborative Research Center SFB/TR TRR 30, which is kindly supported by the Deutsche Forschungsgemeinschaft (DFG).

8 Appendix A: sample results for four infill criteria

The research questions addressed in the actual benchmark 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. This section aims at illustrating some differences between the infill criteria and provide some insight on the performance differences observed in section 5. We particularly focus on the effects of the consideration of the current accuracy of an observation (e. g., by using a Kriging quantile rather than the Kriging mean) and the use of replications (Forrester's approach does not perform any replications while other infill criteria may do). The runs results shown here are somehow typical, but there are of course many situations for which the above comments do not apply.

We show here four runs of the AKG, AEI, EQ90 and RI criteria on the Goldsteinprice with 50% noise, large initial DOE and small budget. Out of the four, EQ90 (Figure 5 C) is clearly less exploratory, since almost all the added observations form a single cluster. This cluster reduces a lot the kriging uncertainty locally, but does not allow global exploration or accurate convergence. Inversely, RI ((Figure 5 D)) does not allow repetitions which results here in very good exploration, but prevents from identifying the valley of the global optimum. The two other criteria, on this example, offer the right trade-off, with a relatively wide exploration and accurate identification of the optimum region.

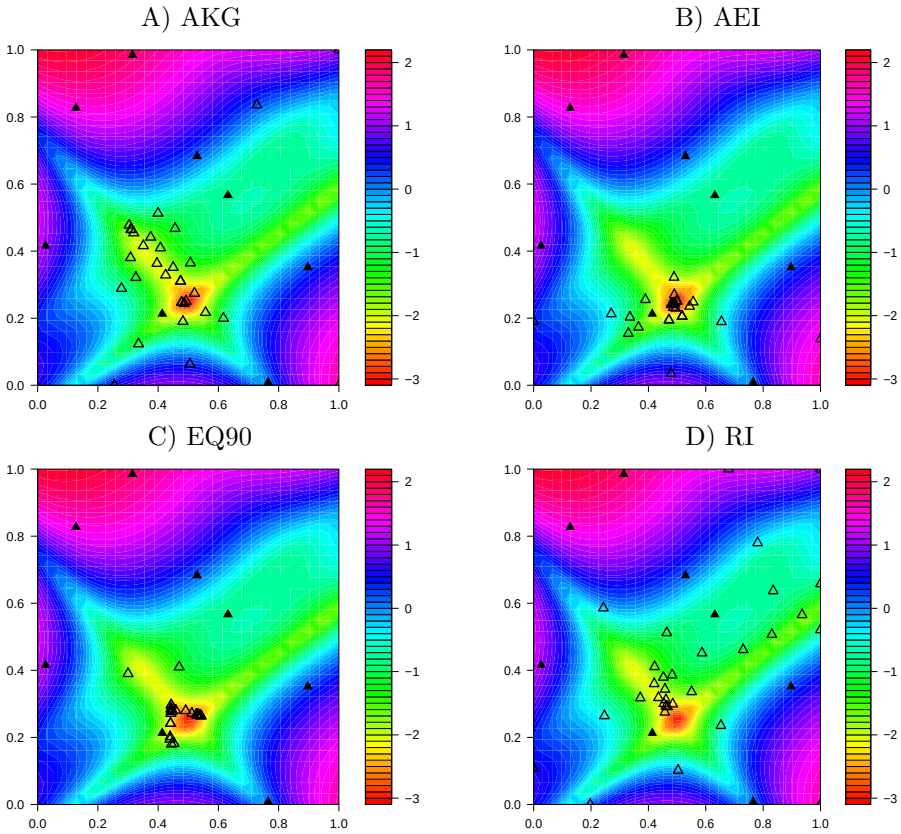


Figure 5: Sample of optimization results obtained on the Goldsteinprice with 50% noise, large initial DOE and small budget. Initial observations are represented with filled triangles. The contour lines represent the actual function.

9 Appendix B: ANOVA tables of the linear models

Table 7: ANOVA table for the Branin test function.

Source	Sum Sq.	d.f.	F	p-val
budget	89.6	1	308.79	**
noise	2312.42	2	3984.63	**
*noise*budget*	1.76	2	3.03	-
init	1.41	1	4.87	*
*init*budget*	0.23	1	0.8	-
init*noise	8.56	2	14.75	**
cov	4.99	1	17.19	**
*cov*budget*	0.48	1	1.65	-
cov*noise	32.82	2	56.55	**
*cov*init*	0.94	1	3.25	-
meth	143.31	10	49.39	**
meth*budget	11.79	10	4.06	**
meth*noise	78.19	20	13.47	**
meth*init	13.64	10	4.7	**
*meth*cov*	5.17	10	1.78	-
lhs	62.92	39	5.56	**
Error	3031.09	10446		
Total	5799.32	10559		

Significance codes: 0 '**' 0.01 '*' 0.05 '-' 1

Table 8: ANOVA table for the Goldstein-Price test function.

Source	Sum Sq.	d.f.	F	p-val
budget	182.73	1	467.59	**
noise	1194.92	2	1528.86	**
noise*budget	7.33	2	9.38	**
init	0.41	1	1.04	-
*init*budget*	0.1	1	0.26	-
init*noise	15.71	2	20.1	**
cov	75.85	1	194.09	**
cov*budget	6.39	1	16.35	**
cov*noise	2.6	2	3.32	*
cov*init	7.03	1	17.98	**
meth	913.87	10	233.85	**
meth*budget	53.03	10	13.57	**
meth*noise	129.96	20	16.63	**
meth*init	19.33	10	4.95	**
meth*cov	64.51	10	16.51	**
lhs	545.58	39	35.8	**
Error	4082.16	10446		
Total	7301.5	10559		

Table 9: ANOVA table for the Hartman4 test function.

Source	Sum Sq.	d.f.	F	p-val
budget	78.29	1	420.45	**
noise	1380.46	2	3707.03	**
*noise*budget*	0.07	2	0.2	-
init	1.01	1	5.44	*
*init*budget*	0.48	1	2.58	-
*init*noise*	0.12	2	0.32	-
cov	0.85	1	4.56	*
*cov*budget*	0.15	1	0.8	-
cov*noise	1.77	2	4.76	*
*cov*init*	0	1	0.01	-
meth	1328.9	10	713.72	**
meth*budget	12.57	10	6.75	**
meth*noise	136.24	20	36.58	**
meth*init	19.18	10	10.3	**
meth*cov	32.18	10	17.28	**
lhs	114.72	39	15.8	**
Error	1944.99	10446		
Total	5051.97	10559		

Table 10: ANOVA table for the Rosenbrock test function.

Source	Sum Sq.	d.f.	F	p-val
budget	11.74	1	52.66	**
noise	284.81	2	638.6	**
noise*budget	2.02	2	4.53	*
init	1.2	1	5.39	*
*init*budget*	0.01	1	0.03	-
init*noise	24.5	2	54.93	**
cov	0.02	1	0.11	-
cov*budget	2.29	1	10.26	**
cov*noise	49.68	2	111.39	**
cov*init	2.73	1	12.25	**
meth	174.23	10	78.13	**
meth*budget	5.53	10	2.48	*
meth*noise	29.91	20	6.71	**
meth*init	5.43	10	2.44	*
meth*cov	21.12	10	9.47	**
lhs	34.3	39	3.94	**
Error	2329.39	10446		
Total	2978.91	10559		

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] Beyer, H., Sendhoff, B.: Robust optimization-a comprehensive survey. Computer methods in applied mechanics and engineering **196**(33-34), 3190–3218 (2007)

[4] 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 Com-

Table 11: ANOVA table for the Hartman6 test function.

Source	Sum Sq.	d.f.	F	p-val
budget	29.8	1	443.53	**
noise	789.33	2	5875.06	**
*noise*budget*	0.39	2	2.92	-
init	0.08	1	1.26	-
init*budget	0.61	1	9.09	**
init*noise	3.8	2	28.25	**
cov	2.71	1	40.33	**
*cov*budget*	0.01	1	0.08	-
*cov*noise*	0.04	2	0.28	-
cov*init	0.32	1	4.8	*
meth	435.35	10	648.07	**
meth*budget	3.46	10	5.15	**
meth*noise	43.55	20	32.41	**
meth*init	13.5	10	20.09	**
meth*cov	1.53	10	2.28	*
lhs	13.58	39	5.19	**
Error	701.73	10446		
Total	2039.79	10559		

Table 12: ANOVA table for the Sphere test function.

Source	Sum Sq.	d.f.	F	p-val
budget	31.61	1	360.54	**
noise	1330.73	2	7590.16	**
*noise*budget*	0.24	2	1.37	-
init	3.19	1	36.36	**
*init*budget*	0.02	1	0.26	-
init*noise	1.64	2	9.36	**
cov	0.04	1	0.46	-
*cov*budget*	0.33	1	3.71	-
*cov*noise*	0.11	2	0.62	-
*cov*init*	0.14	1	1.55	-
meth	587.23	10	669.88	**
meth*budget	7.77	10	8.86	**
meth*noise	140.37	20	80.06	**
meth*init	20.25	10	23.1	**
meth*cov	6.6	10	7.53	**
lhs	25.49	39	7.46	**
Error	915.71	10446		
Total	3071.46	10559		

putation (CEC 2008), 1.-6. June, Hong Kong, pp. 2980–2987. IEEE Press, Piscataway, NJ (2008). DOI 10.1109/CEC.2008.4631199

[5] Cox, D., John, S.: Sdo: A statistical method for global optimization. Multidisciplinary design optimization: state of the art pp. 315–329 (1997)

[6] Cressie, N.: Statistics for spatial data. Terra Nova **4**(5), 613–617 (1992)

[7] Dixon, L., Szegö, G.: Towards global optimisation 2, vol. 2. North Holland (1978)

[8] Emmerich, M.: Single- and multi-objective evolutionary design optimization assisted by gaussian random field metamodels. Ph.D. thesis, Universität Dortmund (2005)

[9] 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)

[10] Forrester, A., Keane, A., Bressloff, N.: Design and Analysis of "Noisy" Computer Experiments. AIAA journal **44**(10), 2331 (2006)

[11] 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)

[12] Gramacy, R., Lee, H.: Optimization under unknown constraints. Arxiv preprint arXiv:1004.4027 (2010)

[13] Hardy, M.: Regression with dummy variables. 91-93. Sage Publications, Inc (1993)

[14] Huang, D., Allen, T., Notz, W., Miller, R.: Sequential kriging optimization using multiple-fidelity evaluations. Structural and Multidisciplinary Optimization **32**, 369–382 (2006)

[15] 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)

[16] Humphrey, D., Wilson, J.: A revised simplex search procedure for stochastic simulation response surface optimization. INFORMS Journal on Computing **12**(4), 272–283 (2000)

[17] 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)

[18] Jin, R., Chen, W., Simpson, T.: Comparative studies of metamodelling techniques under multiple modelling criteria. Structural and Multidisciplinary Optimization **23**, 1–13 (2001)

[19] Jones, D.: A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization **21**(4), 345–383 (2001)

[20] Jones, D., Schonlau, M., Welch, W.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization **13**(4), 455–492 (1998)

[21] Kleijnen, J.: Design and analysis of simulation experiments, vol. 111. Springer Verlag (2007)

[22] 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)

[23] 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)

[24] Matheron, G.: Le krigeage universel. Cahiers du centre de morphologie mathématique **1** (1969)

[25] Osborne, M., Garnett, R., Roberts, S.: Gaussian processes for global optimization. In: 3rd International Conference on Learning and Intelligent Optimization (LION3), pp. 1–15 (2009)

[26] 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)

[27] 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)

[28] Rasmussen, C., Williams, C.: Gaussian processes for machine learning. MIT Press (2006)

[29] 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)

[30] Sacks, J., Welch, W., Mitchell, T., Wynn, H.: Design and analysis of computer experiments. Statistical science pp. 409–423 (1989)

[31] 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)

[32] 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)

[33] Santner, T., Williams, B., Notz, W.: The design and analysis of computer experiments. Springer (2003)

[34] Sasena, M.: Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations. Ph.D. thesis, University of Michigan (2002)

[35] Scott, W., Frazier, P., Powell, W.: The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression. SIAM Journal on Optimization **21**, 996 (2011)

[36] Sekhon, J., Mebane, W.: Genetic optimization using derivatives. Political Analysis **7**(1), 187 (1998)

[37] 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)

[38] Srinivas, N., Krause, A., Kakade, S., Seeger, M.: Gaussian process optimization in the bandit setting: No regret and experimental design. In: 27th International Conference on Machine Learning (ICML 2010) (2010)

[39] 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)

[40] Villemonteix, J., Vazquez, E., Walter, E.: An informational approach to the global optimization of expensive-to-evaluate functions. Journal of Global Optimization **44**(4), 509–534 (2009)

[41] Wagner, T., Wessing, S.: On the effect of response transformations in sequential parameter optimization. Evolutionary Computation **20**(2), 229–248 (2012). DOI 10.1162/EVCO_a_00061

[42] Yin, J., Ng, S., Ng, K.: Kriging metamodel with modified nugget-effect: The heteroscedastic variance case. Computers & Industrial Engineering **61**(3), 760–777 (2011)