



# Efficient global optimization with ensemble and selection of kernel functions for engineering design

Pramudita Satria Palar<sup>1</sup> · Koji Shimoyama<sup>1</sup>

Received: 2 April 2018 / Revised: 22 June 2018 / Accepted: 18 July 2018 / Published online: 28 July 2018  
© Springer-Verlag GmbH Germany, part of Springer Nature 2018, corrected publication September/2018

## Abstract

In this paper, we investigate the use of multiple kernel functions for assisting single-objective Kriging-based efficient global optimization (EGO). The primary objective is to improve the robustness of EGO in terms of the choice of kernel function for solving a variety of black-box optimization problems in engineering design. Specifically, three widely used kernel functions are studied, that is, Gaussian, Matérn-3/2, and Matérn-5/2 function. We investigate both model selection and ensemble techniques based on Akaike information criterion (AIC) and cross-validation error on a set of synthetic (noiseless and noisy) and non-algebraic (aerodynamic and parameter tuning) optimization problems; in addition, the use of cross-validation-based local (i.e., pointwise) ensemble is also studied. Since all the constituent surrogate models in the ensemble scheme are Kriging models, it is possible to perform EGO since the Kriging uncertainty structure is still preserved. Through analyses of empirical experiments, it is revealed that the ensemble techniques improve the robustness and performance of EGO. It is also revealed that the use of Matérn-kernels yields better results than those of the Gaussian kernel when EGO with a single kernel is considered. Furthermore, we observe that model selection methods do not yield any substantial improvement over single kernel EGO. When averaged across all types of problem (i.e., noise level, dimensionality, and synthetic/non-algebraic), the local ensemble technique achieves the best performance.

**Keywords** Efficient global optimization · Kernel function · Surrogate model · Model selection · Model ensemble

## 1 Introduction

Numerical optimization plays an important role in engineering design, in order to discover more efficient designs in terms of performance. In engineering design, an optimization method is usually coupled with partial differential equation (PDE) solvers in order to evaluate and optimize the performance of new designs. One challenge with the majority of PDE-based engineering optimization problems is that they are typically computationally expensive and provide no gradient information. Nowadays, computer

experiments take great advantage of surrogate models in order to accelerate the optimization process. One merit of surrogate models is that they are not bounded by any discipline, thus can be used in a variety of disciplines.

Among various surrogate models, Kriging models (or also known as Gaussian processes) (Kriging 1951; Matheron 1969) have been widely applied in engineering design optimization. The use of Kriging models for modeling computer experiments was firstly initiated by Sacks et al. (1989) and their use has particularly flourished in the field of optimization. Note that direct optimization using Kriging prediction might lead to the discovery of false/local optima. To handle this problem, an efficient optimization strategy based on Kriging models called efficient global optimization (EGO) that uses the expected improvement (EI) metric (Moćkus 1975) has been developed in order to improve the efficiency and robustness of surrogate-based optimization (Jones et al. 1998). In the machine learning community, this optimization strategy is popular under the name of Bayesian optimization. The EGO framework is not just limited to the EI metric; implementation of other metrics such as the probability of improvement (Jones 2001), entropy search

---

Responsible Editor: Erdem Acar

✉ Pramudita Satria Palar  
palar@edge.ifs.tohoku.ac.jp  
Koji Shimoyama  
shimoyama@edge.ifs.tohoku.ac.jp

<sup>1</sup> Institute of Fluid Science, Tohoku University,  
Sendai, Miyagi Prefecture 980-8577, Japan

(Hennig and Schuler 2012), and bootstrapped EI (Kleijnen et al. 2012) are also possible. The EGO framework has been successfully employed in various engineering applications such as aerodynamic design (Jeong et al. 2005), gait parameter optimization in robotics (Tesch et al. 2011), and an injection molding process (Shi et al. 2010), to name a few. Implementation of EGO with various enrichment criteria for solving noisy problems has also been studied by Picheny et al. (2013).

For efficient surrogate modeling using Kriging, it is important to choose the most appropriate form for a specific application by selecting a proper kernel function for constructing the covariance matrix. To this end, several studies have been performed within various contexts such as optimization and uncertainty quantification. Mukhopadhyay et al. (2017) compared the effect of various kernel functions on Kriging approximation in the context of uncertainty quantification in the dynamics of a composite shell and argued that the Gaussian correlation function is the best in terms of approximation error. Similarly, experiments by Acar (2013) showed that Gaussian is superior to exponential and linear correlation functions. The use of advanced correlation functions such as Matérn kernels for EGO in the context of engineering design optimization is still rare. Matérn kernels are recommended and suggested for real-world processes since they address the problem of unrealistically smooth behavior due to the Gaussian correlation function (Rasmussen and Williams 2006). Nevertheless, presently, no comprehensive comparison between Matérn and Gaussian correlation function within the context of EGO has been done.

The methodology of the ensemble of surrogates has recently attracted renewed interest. Instead of choosing the most potential surrogate model for a specific application, the main reason for utilizing the ensemble of surrogates is to make the best use of an individual surrogate model to create a robust and/or strong single learner. Firstly initiated for a function approximation problem by Goel et al. (2007), the technique was then improved by using an optimization formulation to minimize the error metric (Acar and Rais-Rohani 2009). A comprehensive study on the ensemble of surrogates was then performed by Viana et al. (2009). It is also possible to construct a local ensemble of surrogates that uses a non-constant weight function to average the surrogate (Acar 2010). Zhang et al. (2012) proposed a hybrid surrogate model that exploits the local measure of accuracy. Recently, Liu et al. proposes a local ensemble of surrogates with multiple radial basis functions and applied the method for blade design (Liu et al. 2016). Ben Salem and Tomaso (2018) discusses an automatic selection technique for general surrogate models that aims for good accuracy. Ensemble of surrogates techniques have been applied to various interesting engineering cases such

as robust parameter design (Zhou et al. 2013) and axial compressor blade shape optimization (Samad et al. 2008).

There are also similar developments in multiple kernel techniques within the machine learning community. In particular, Archambeau and Bach (2011) discuss the convex combination of kernel matrices to improve the generalization on unseen data. Durrande et al. (2011) proposes an additive kernel for a Gaussian process model, where multiple kernels with various length scales are added. In other related literature, although not in the context of Kriging/Gaussian processes, is the hierarchical kernel learning technique (Bach 2009). The main difference between these techniques and ensemble of surrogates is that the kernels are typically combined inside the surrogate model itself. In this paper, we focus on the ensemble of Kriging models, that is, multiple Kriging models are constructed and then combined into one model through weighted sum formulation. The ensemble formulation is also convenient for general purposes since one can use various implementations of Kriging/Gaussian processes, e.g., ooDACE (Couckuyt et al. 2014) and UQLab (Marelli and Sudret 2014), with little modification to utilize the ensemble of surrogates or model selection.

One aspect that is often overlooked is that the goal of optimization is not to pursue global accuracy but to efficiently discover the global optimum. Therefore, it is more important that the surrogate model should be more accurate in the vicinity of the global optimum, while still useful enough to guide the optimization process into the near-global optimum location. Furthermore, one disadvantage of creating an ensemble of surrogate models that do not provide uncertainty structure is that it becomes non-trivial to perform Bayesian optimization since the uncertainty structure vanishes after the ensemble process. In a relevant past study, Ginsbourger et al. (2008) suggest the mixture of Kriging with multiple kernel functions to assist EGO; they propose mixing Kriging models via Akaike Information Criterion (AIC)-based weight determination. However, this study is relatively limited since no extensive empirical experiments were performed.

In this paper, we revisit the concept of mixing a set of Kriging models with multiple kernel functions for more effective EGO. Besides the mixture of Kriging models, Kriging model selection (i.e., choosing the model with greatest potential) is also investigated. The primary objective of this research is to develop a robust and powerful EGO method that uses information from multiple kernel functions. In this paper, we empirically investigate several approaches to analyze the best ensemble or model selection method in terms of optimization performance. To this end, empirical tests on selected synthetic and non-algebraic problems were performed. Furthermore, we also studied the impact of deterministic noise as in typical

computer experiments on the performance of EGO with single and multiple kernel functions. Specifically, we studied the global ensemble approach of Acar and Rais-Rohani (2009), local ensemble approach of Liu et al. (2016), AIC-based mixtures suggested by Ginsbourger et al. (2008), and model selection based on AIC and cross-validation (CV) error. The engineering problems selected to represent general engineering problems come from the field of aerodynamic optimization. EGO itself has been extensively used in the field of aerodynamic design and optimization (Jeong et al. 2005; Kanazaki et al. 2015; Bartoli et al. 2016; Namura et al. 2016).

The remainder of this paper is structured as follows: in Section 2 we explain the basic of Kriging model and the choice of kernel functions; in Section 3 we detail the techniques for the ensemble and model selection of Kriging with various kernel functions; in Section 4 we present and discuss the results from the computational studies; the paper is then concluded and recommendations for the future work are given in Section 5.

## 2 Kriging surrogate model

### 2.1 Basics

Kriging approximates the relationship between the input  $\mathbf{x} = \{x_1, \dots, x_m\}^T$ , where  $m$  is the dimensionality of the decision variables, and the output  $y$  is a mean of the realization of a random field that is governed by the prior covariances. To construct a Kriging model, a finite set of experimental design (ED)  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}^T$  and the corresponding responses  $\mathbf{y} = \{y_1, \dots, y_n\}^T$ , where  $n$  is the number of sampling points, in the design space need to be collected first.

Within the context of ordinary Kriging (OK), the black box function is approximated using the following expression:

$$Y(\mathbf{x}) = \mu_{KR} + Z(\mathbf{x}), \tag{1}$$

where  $\mu_{KR}$  is the mean of the Kriging approximation and  $Z(\mathbf{x})$  denotes the deviation from the mean.

The main assumption of Kriging models are that the difference between the responses of two different points is small when their distance is also small. This assumption is modeled by the use of a kernel (or correlation) function. The correlation between two points  $\mathbf{x}^{(i)}$  and  $\mathbf{x}^{(j)}$  is described by  $\mathbf{R}_{ij} = \text{corr}[Z(\mathbf{x}^{(i)}), Z(\mathbf{x}^{(j)})]$ , that is further described by the hyperparameters  $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$ .

The OK prediction for an arbitrary input variable reads as follows:

$$\hat{y}(\mathbf{x}) = \mu_{KR} + \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\mu_{KR}), \tag{2}$$

with the mean-squared error of the Kriging prediction  $\hat{\sigma}^2(\mathbf{x})$  reads as:

$$\hat{\sigma}^2(\mathbf{x}) = \sigma^2 \left( 1 - (\mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x})) + \left( 1 - \mathbf{1}^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}) \right)^2 \times \left( \mathbf{1}^T \mathbf{R}^{-1} \mathbf{1} \right)^{-1} \right). \tag{3}$$

Here,  $\mathbf{R}$  is the  $n \times n$  matrix with the  $(i, j)$  entry is  $\text{corr}[Z(\mathbf{x}^{(i)}), Z(\mathbf{x}^{(j)})]$ ,  $\mathbf{r}(\mathbf{x})$  is the correlation vector between  $\mathbf{x}$  and  $\mathcal{X}$  whose  $(i, 1)$  entry is  $\text{corr}[Z(\mathbf{x}^{(i)}), Z(\mathbf{x})]$ ,  $\mathbf{1}$  is a vector of ones with length  $n$ , and  $\sigma^2$  is the process variance. The Kriging mean term, i.e.,  $\mu_{KR}$ , is obtained as follows:

$$\mu_{KR} = (\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \mathbf{R}^{-1} \mathbf{y}. \tag{4}$$

Readers are referred to Jones et al. (1998) and Sacks et al. (1989) for more information regarding Kriging model.

### 2.2 Kernel function

To construct the Kriging model, it is necessary to determine the type of kernel function that models the correlation between design points. In this paper, we opt for the Gaussian and Matérn kernel functions and study their performance individually and as a mixture of models on the optimization of synthetic and non-algebraic problems. The Gaussian kernel function is selected primarily due to its wide use in engineering design optimization. The Gaussian kernel assumes that the response function is smooth, however, this assumption is somehow unrealistic for real-world processes (Stein 2012). Instead, Stein recommends the use of Matérn class correlation function (Stein 2012).

#### 2.2.1 Gaussian

The Gaussian kernel function is defined as:

$$R(\theta)^{(i)} = \exp\left(-\frac{h^2}{2\theta^2}\right). \tag{5}$$

where  $h = \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|$ , with the indices  $i$  and  $j$  indicating two distinct design points, and  $\theta$  is the length scale of the kernel. In multivariable approximation, there are  $m$  length scales (i.e., hyperparameters) to be tuned, that is,  $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$ . Considering its wide use, including Gaussian as a constituent of the ensemble model in our study is then relevant for our investigation.

#### 2.2.2 Matérn class

The use of Matérn class functions for Kriging model was proposed by Stein (2012) based on the work of Matérn.

The general form of Matérn kernel function is expressed as

$$R(h, \theta, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\sqrt{\nu}\frac{|h|}{\theta}\right)^\nu \mathcal{K}_\nu \left(2\sqrt{\nu}\frac{|h|}{\theta}\right), \quad (6)$$

where  $\nu \geq 1/2$  is the shape parameter,  $\Gamma$  is the Gamma function, and  $\mathcal{K}_\nu$  is the modified Bessel function of the second kind. A specific Matérn kernel function can be constructed by using a specific value of  $\nu$ ; however, a half-integer value of  $\nu$  (i.e.,  $\nu = p + 1/2$ , where  $p$  is a non-negative integer) is typically used for a simple expression of Matérn kernel function. According to Rasmussen and Williams (2006), the most interesting cases for machine learning cases are  $\nu = 3/2$  and  $\nu = 5/2$ .

For  $\nu = 3/2$  the formulation of Matérn kernel function is

$$R(h, \theta, \nu = 3/2) = \left(1 + \frac{\sqrt{3}|h|}{\theta}\right) \exp\left(-\frac{\sqrt{3}|h|}{\theta}\right), \quad (7)$$

while for  $\nu = 5/2$ , the Matérn kernel function is defined as

$$R(h, \theta, \nu = 5/2) = \left(1 + \frac{\sqrt{5}|h|}{\theta} + \frac{5h^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}|h|}{\theta}\right). \quad (8)$$

Matérn-3/2 and Matérn-5/2 are two forms of Matérn kernel function that are widely used to model real-world processes. We, therefore, used these two forms of Matérn kernel function in our study and compare them with the standard Gaussian.

### 2.3 Hyperparameter estimation method

Every kernel function possesses a set of hyperparameters that need to be optimized. The standard technique to construct Kriging is to find the set of hyperparameters that optimize the so-called likelihood function, a method commonly termed maximum likelihood (ML) estimation, defined as (Jones 2001)

$$\mathcal{L}(\mu_{KR}, \sigma^2, \theta) = \frac{1}{\sqrt{(2\pi\sigma^2)^{n/2} |\mathbf{R}(\theta)|}^{\frac{1}{2}}} \times \exp\left(-\frac{1}{2} \frac{(\mathbf{y} - \mathbf{1}\mu_{KR})^T \mathbf{R}(\theta)^{-1} (\mathbf{y} - \mathbf{1}\mu_{KR})}{\sigma^2}\right), \quad (9)$$

where the maximum likelihood estimates of the Kriging variance  $\hat{\sigma}^2$  is computed as follows:

$$\hat{\sigma}^2(\theta) = \frac{1}{n} (\mathbf{y} - \mathbf{1}\mu_{KR})^T \mathbf{R}(\theta)^{-1} (\mathbf{y} - \mathbf{1}\mu_{KR}). \quad (10)$$

For this study, we use a hybrid of genetic algorithm with local hill climbing to optimize the hyperparameters. The local hill climbing is executed on the hyperparameters found by GA in order to further refine the optimized hyperparameters.

### 2.4 Expected improvement

EGO utilizes a special metric that take into account both the Kriging prediction and estimation error in order to drive the optimization process. The most common metric to be used for EGO is the EI metric, defined as

$$E[I(\mathbf{x})] = (y_{min} - \hat{y}(\mathbf{x}))\Phi\left(\frac{y_{min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) + s(\mathbf{x})\varphi\left(\frac{y_{min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right), \quad (11)$$

where  $y_{min}$  is the function value of the best solution observed, and  $\Phi(\cdot)$  and  $\varphi(\cdot)$  are the cumulative distribution function and probability density function of a normal distribution, respectively. The next sample to be added into the experimental design is then found by maximizing the EI metric. We maximize the EI value by using the same strategy with that of hyperparameters optimization.

### 2.5 Handling noisy problems via reinterpolation

To handle noisy problems, we use the reinterpolation procedure of Forrester et al. (2006). In noisy problems, we use regressing Kriging where a regression factor term  $\lambda$  is added to the correlation matrix  $\mathbf{R}$ , that is,  $\mathbf{R} + \lambda \mathbf{I}$ , where  $\mathbf{I}$  is an identity matrix. Here,  $\lambda$  is also tuned in the range of  $10^{-6}$  and  $10^{-1}$  in a logarithmic scale to improve the likelihood of the data (Forrester et al. 2006).

The primary goal of the reinterpolation procedure is to avoid resampling that is caused by the non-zero error at sampling points due to the use of regression Kriging. The basic idea of reinterpolation is to construct an interpolating model through the Kriging predictor from the Kriging regression. The classical EI criterion can then be employed on the new interpolating model to seek the next point to be evaluated as in the noise-free problems. In this paper, we employ Kriging regression and reinterpolation on noisy synthetic problems and non-algebraic problems. Kriging with reinterpolation is also suggested by Picheny et al. (2013) in cases where the noise level is difficult to be identified beforehand; in this paper, we assume that we have no prior knowledge regarding noise level on the noisy synthetic problems.

### 3 Utilizing multiple kernel functions for efficient global optimization

In practice, it is not easy to select the best kernel function for a specific application. To handle this problem, we advocate the use of Kriging models with multiple kernel functions in order to boost the robustness and performance of EGO. This can be done via two ways: creating an ensemble of

surrogates or selecting a Kriging model with a specific kernel function in an online fashion. By these methods, we seek to improve the robustness of EGO in the situation with misspecified kernel function. Furthermore, we expect that an improvement in optimization performance can be achieved through the combination of multiple learners, which are the Kriging surrogate models in this case.

### 3.1 Ensemble of Kriging models

The general form of the ensemble of surrogate models, which can also be interpreted as a mixture of experts, reads as

$$\hat{y}_{ens}(\mathbf{x}) = \sum_{i=1}^K w_i(\mathbf{x}) \hat{y}_i(\mathbf{x}) \tag{12}$$

where  $K$  is the number of surrogate models,  $\{\hat{y}_1(\mathbf{x}), \dots, \hat{y}_K(\mathbf{x})\}$  is the vector of function predictions of the  $K$  surrogate models, and  $\mathbf{w}(\mathbf{x}) = \{w_1(\mathbf{x}), \dots, w_K(\mathbf{x})\}$  is the weight vector that defines the contribution of individual surrogate models. There are various existing techniques to compute the weight function; in this paper, we use several proven methods to construct the mixture of Kriging with various kernel functions. We briefly explain the techniques that we studied in this paper below.

#### 3.1.1 Global ensemble via cross validation

The global ensemble approach assigns a constant weight for each surrogate model in the given range of the design space. Goel et al. (2007) proposed a heuristic based on the individual and total CV error of the surrogate models. The method is further refined by Acar and Rais-Rohani (2009) that uses Bishop’s technique (Bishop 1995) to compute the weight so as to directly minimize the CV error; we use this approach in this paper.

The computation of CV error is vital to the construction of an error-based ensemble of surrogates; mostly leave-one-out CV (LOOCV) error is employed for this purpose. Fortunately for Kriging models, the LOOCV error can be computed analytically without the need to build  $n$  Kriging models (Dubrule 1983). We first define  $\mathbf{e}_i = \{e_i^{(1)}, \dots, e_i^{(n)}\}$  as the vector of CV error for the surrogate  $i$ , where  $e_i^{(j)} = y(\mathbf{x}^{(j)}) - \hat{y}^{(i,-j)}(\mathbf{x}^{(j)})$  is the cross-validation error for sample  $j$  with sample  $j$  removed from the experimental design and  $\hat{y}^{(i,-j)}(\mathbf{x}^{(j)})$  is the prediction of surrogate  $i$  without taking into account sample  $j$ .

The weight vector  $\mathbf{w}$  is found by solving the following minimization problem

$$\min_{\mathbf{w}} \text{MSE}_{ens} = \mathbb{E} \left( e_{WAS}^2(\mathbf{x}) \right) = \mathbf{w}^T \mathbf{C} \mathbf{w}, \tag{13}$$

where  $\text{MSE}_{ens}$  is the mean squared error of the global ensemble (i.e., the objective function to be minimized),  $\mathbf{C}$  is

an  $K \times K$  matrix of cross validation error for all surrogate models with its element is computed as  $c_{ij} = \frac{1}{n} \mathbf{e}_i^T \mathbf{e}_j$ , and  $e_{WAS}$  is the squared error of the weighted average surrogate. To ensure a positive weight, we follow Viana et al.’s suggestion (Viana et al. 2009) to solely used the diagonal matrix of  $\mathbf{C}$  to compute  $\mathbf{w}$  using Lagrange multipliers.

#### 3.1.2 Global ensemble via Akaike information criterion

The second global ensemble that we investigate is the ensemble via Akaike weights that is based on AIC (Ginsbourger et al. 2008). Central to this method is AIC, which denotes the Kullback-Leibler distance between the truth and the model. AIC is computed as

$$AIC = -2 \ln(\mathcal{L}) + 2N_f, \tag{14}$$

where  $N_f$  is the number of free parameters. In the context of ordinary Kriging, the constant mean and the process variance should also be counted as the free parameters. The total number of free parameters for interpolating ordinary Kriging in the current implementation is then the length of  $\boldsymbol{\theta} + 2$ . Furthermore,  $\lambda$  is also counted as a free parameter in the regression Kriging. After the AIC is computed, the Kriging model with the lowest AIC (i.e.,  $AIC_{min}$ ) is firstly used as the reference model. We can then compute the relative AIC difference between other models and the model with the lowest AIC, that is,  $\Delta AIC_i = AIC_i - AIC_{min}$ . The Akaike weight for a model  $i$  is then computed as Burnham and Anderson (2003).

$$w_i = \frac{\exp(-0.5 \Delta AIC_i)}{\sum_{j=1}^K \exp(-0.5 \Delta AIC_j)}. \tag{15}$$

In practice, it is better to use the corrected version of AIC, i.e.,  $AIC_c$  to take into account the factor of sample size and increase the penalty for model complexity with small data sets. The  $AIC_c$  is defined as

$$AIC_c = AIC + \frac{2N_f^2 + 2N_f}{n - N_f - 1}. \tag{16}$$

Although its use in Kriging with multiple kernels was first mentioned by Ginsbourger et al. (2008), the lack of empirical experiment left the knowledge regarding the potential of this method scarce. Moreover, Martin and Simpson (2005) argued that  $AIC_c$  is a better predictor of surrogate model accuracy compared to CV error.

#### 3.1.3 Local ensemble via cross validation

Another alternative to the global ensemble is the local ensemble (i.e., pointwise) approach that employs a non-constant weight function. Our main reasoning to consider local

ensemble is that there is a good reason to trust a specific kernel function in a certain region, while another might be more accurate in other regions of the design space. In this paper, we opt for Liu et al.'s approach (Liu et al. 2016) to construct local ensemble of Kriging models. To compute the weight for surrogate  $i$  at a certain design point, the following equation is used:

$$w_i(\mathbf{x}) = \begin{cases} \text{if } \mathbf{x} \neq \mathbf{x}_j : \frac{d_j^{-B_j\Theta} W_{ji}}{\sum_{j=1}^n d_j^{-B_j\Theta}}, \\ \text{if } \mathbf{x} = \mathbf{x}_j : W_{ji} \end{cases} \quad (17)$$

where  $\Theta$  is the attenuation coefficient that is automatically selected using cross-validation error,  $d_j$  is the distance between  $\mathbf{x}$  and  $\mathbf{x}_j$ ,  $B_j$  is the normalized global accuracy of the constituent model that yields the lowest error at  $\mathbf{x}_j$ , and  $\mathbf{W}$  is the observed weight matrix (i.e., one at the sampling points and zero elsewhere for the surrogate that yields the lowest CV error). Liu et al.'s method constructs a non-constant weighting function through a 0-1 weight strategy, where the weight is one for a specific surrogate at locations where that surrogate yields the lowest CV error. The value of  $\Theta$ , which is set to a value between 0 and 15 in steps of 1, determines the influence of the weight matrix to unobserved points as a function of distance. In this case,  $\Theta = 0$  results in a situation that mimics that of the global ensemble while  $\Theta = 15$  yields the highest influence of the weight matrix. Readers are referred to Liu et al. (2016) for a detailed explanation of the method.

### 3.1.4 Computation of expected improvement

The prediction of the ensemble of Kriging models is computed similarly with (12). On the other hand, the corresponding mean-squared error and EI are, respectively, computed by Ginsbourger et al. (2008)

$$s_{ens}^2(\mathbf{x}) = \sum_{i=1}^K w_i s_i^2(\mathbf{x}) + \sum_{i=1}^K w_i [(\hat{y}_i(\mathbf{x}) - \hat{y}_{ens}(\mathbf{x}))^2], \quad (18)$$

and

$$EI_{ens}(\mathbf{x}) = \sum_{i=1}^K w_i EI_i(\mathbf{x}). \quad (19)$$

The MATLAB-based UQLab tool is employed to construct the Kriging model (Marelli and Sudret 2014). We also modified the source code of UQLab so that we could tune the nugget term for noisy optimization.

### 3.2 Online selection of kernel function

In contrast to the ensemble approach, the online selection approach chooses a Kriging model with a specific kernel

function at each iteration of EGO. To this end, we simply select the Kriging model via one of the following two criteria: the lowest CV error or the lowest AIC. The implementation of this online selection technique is fairly simple: one needs to build a number of Kriging surrogate models using various kernel functions at each iteration, the LOOCV error/AIC of these models are then assessed and the model that yields the lowest LOOCV error/AIC is selected.

### 3.3 Research questions

With the primary goal to improving the robustness of EGO and, if possible, boosting its search performance, we have specific questions to be answered regarding its performance aspects:

a. *Could model selection and the mixture of Kriging with multiple kernel functions improve the performance and robustness of EGO?*

We expect that the performance of the ensemble methods to be robust in solving a wide variety of problems compared to that of model selection. We also hypothesize that model selection would yield better performance than that of a single Kernel function. Comparison with Kriging of single kernel functions is also of our interest since we want to demonstrate the advantage of utilizing multiple kernel functions. Furthermore, we hypothesized that using local ensemble would produce superior solutions versus those of the global ensemble since the former has more control on the locality of the objective function.

b. *What is the effect of kernel function selection on the performance of EGO?*

In this paper, we wanted to further investigate the effect of kernel function on the performance of EGO. Three kernel functions are selected for this purpose- that is, Gaussian, Matérn-3/2, and Matérn-5/2. The Gaussian function is selected mainly due to its popularity for engineering design optimization, besides, it also has remarkable approximation power compared to that of exponential and linear kernel (Acar 2013). On the other hand, Matérn-3/2 and Matérn-5/2 are selected as the other kernel functions of interest due to their wide use within the machine learning context.

c. *To what extent does numerical noise affects the choice of kernel, model selection, and mixture of Kriging?*

We wanted to answer this question specifically in the context of EGO. It is well known that computer simulation-based optimization might be subject to numerical noise due

to truncation errors or changing meshing schemes. This numerical noise, in turn, might affect the performance of the kernel, model selection, and the mixture of Kriging. Depending on the numerical scheme, numerical noise can be suppressed into a near noise-free level; or if it is just too difficult to suppress the noise, we have to take its impact into account when performing EGO. It is then of utmost importance to analyze the performance in both the noise-free and noisy setting so that rich discussion could be made concerning this issue. In this paper, we varied the level of noise from low to moderate in order to gain further insight and shed light on this question.

## 4 Results and discussions

### 4.1 Experimental setup

#### 4.1.1 Test problems

We numerically investigate and compare the performance of EGO with kernel selection, model mixture, and single kernel function on seven synthetic and five non-algebraic problems. In this paper, we select seven synthetic problems for our study, that is, Branin (F1), Sasena (F2), Hosaki (F3), Hartmann-3 (F4), Hartmann-6 (F5), modified Rosenbrock function (F6) from Sobester et al. (2004), and high-dimensional sphere function (F7). In general, we set the initial sample size (i.e.,  $n_{int}$ ) to  $10 \times m$  except for the Hosaki, modified Rosenbrock, sphere, and non-algebraic problems. For all synthetic functions, we vary the noise level from 0% (i.e., noiseless), 1% (low), and 5% (moderate) of the function standard deviation added to the objective function. Note that although our main concern is to solve problems with deterministic noise, the simulated noise does not differentiate repeatable and non-repeatable noise.

Study of non-algebraic problems is necessary in order to gain insight into the performance of various considered optimization methods on solving engineering problems. In this article, studies on several aerodynamic optimization cases as representatives for general engineering problems were performed. The problems are the shape optimization of an airfoil in a subsonic and transonic flow. Here, we use XFOIL (Drela 1989) and the inviscid solver from SU2 CFD (Palacios et al. 2013) for the subsonic and transonic case, respectively. The airfoil is parameterized using nine-variable PARSEC (Sobieczky 1999) and 16-variable class shape transformation (CST) (Kulfan 2008) for the subsonic and transonic case, respectively. We consider two subcases by varying the design condition: (1) optimization with fixed Mach number ( $M$ ) and angle of attack ( $\alpha$ ), and (2) fixed  $M$  and lift coefficient ( $C_l$ ), with the lift-to-drag ratio (i.e.,  $L/D$ )

as the objective function for all cases. In total we have four different aerodynamic optimization problems as follows:

1. Subsonic airfoil, flight condition:  $M = 0.3$ ,  $C_l = 0.5$ ,  $Re = 3 \times 10^6$ .
2. Subsonic airfoil, flight condition:  $M = 0.3$ ,  $\alpha = 2^\circ$ ,  $Re = 3 \times 10^6$ .
3. Transonic airfoil, flight condition:  $M = 0.73$ ,  $C_l = 0.7$ .
4. Transonic airfoil, flight condition:  $M = 0.73$ ,  $\alpha = 2^\circ$ .

where  $Re$  is the Reynolds number.

Besides tests on aerodynamic problems, we also performed a test on the parameter optimization of SVR model applied to the Abalone data set. Although this is clearly not an engineering problem, it serves as a good test problem in the sense that it features several characteristics that are suitable for this study, that is, deterministic noise and non-linear response surface. The naming, number of maximum update ( $n_{upd}$ ),  $n_{int}$  and other details of the test problems are detailed in Table 1, Appendices A and B.

Note that the aerodynamic cases and SVR problem considered in this paper are relatively cheap, especially the subsonic airfoil cases, where the use of surrogate-based optimization for solving such problems is actually overkill. However, our aim is to capture the complexity of real-world problems in terms of the input-output relationship. That is to say that the use of a high-fidelity solver does not necessarily translate into a complex response surface. Hence, tests on low-fidelity problems are really beneficial to assess the performance of optimizers that involve randomness. In this paper, the experiment was repeated in 20 independent runs for all problems.

#### 4.1.2 Comparison methodologies

The performance is assessed using the obtained best solution at the end of the search and the convergence rate of the best solution observed. To that end, we use the log of optimality gap (i.e.,  $OG$ ), reads as

$$\log_{10}(OG) = \log_{10}(y_{min} - y_{opt}) \quad (20)$$

where  $y_{min}$  and  $y_{opt}$  is the observed minimum of the objective function and the true optimum of the problem, respectively. The true optimum solutions are available, or can be easily discovered, for synthetic problems. On the other hand, we use the best objective function observed for non-algebraic problems since the true optimum solutions are unknown beforehand. For noisy problems,  $OG$  is measured by using the objective function value of the corresponding noiseless functions rather than those with the noise. This allows us to assess the performance of the optimizer in locating the true optimum of the problem (i.e., noiseless case) in the presence of noise.

**Table 1** Test problems employed in this paper and parameter settings. For synthetic problems, the noise level is varied by 0%, 1%, and 5%

No.	Problem	$m$	$n_{int}$	$n_{upd}$	Design variables
F1	Branin	2	20	40	$[0, 1]^2$
F2	Sasena	2	20	40	$[0, 5]^2$
F3	Hosaki	2	12	40	$[0, 5]^2$
F4	Hartmann-3	3	30	60	$(0, 1)^3$
F5	Hartmann-6	6	60	80	$(0, 1)^6$
F6	Mod. Rosenbrock	5	32	64	$[-1, 1]^5$
F7	Sphere	9	30	80	$[-5.12, 5.12]^9$
A1	SVR parameters tuning	2	15	40	$[10^{-3}, 10^3]^2$
A2	Subsonic airfoil, fixed $C_l$	9	30	80	See Appendix A
A3	Subsonic airfoil, fixed $\alpha$	9	30	80	See Appendix A
A4	Transonic airfoil, fixed $C_l$	16	30	60	See Appendix A
A5	Transonic airfoil, fixed $\alpha$	16	30	60	See Appendix A

To measure the convergence rate we compute the area under a curve (AUC) of the optimality gap computed using trapezoidal rule reads as

$$AUC = \frac{1}{2}(OG_0 + 2OG_1 + 2OG_2 + \dots + 2OG_{n_{upd}-1} + OG_{n_{upd}}) \tag{21}$$

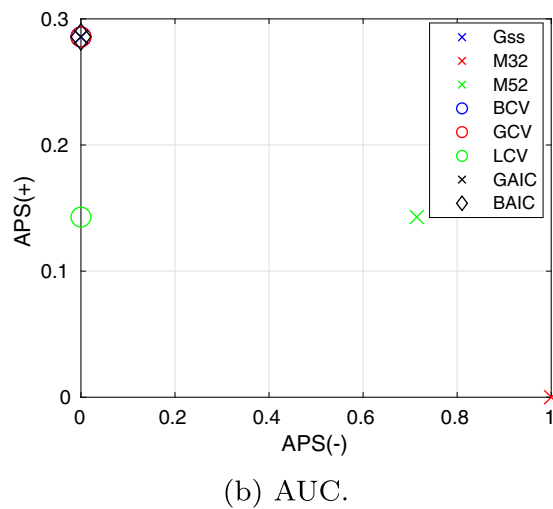
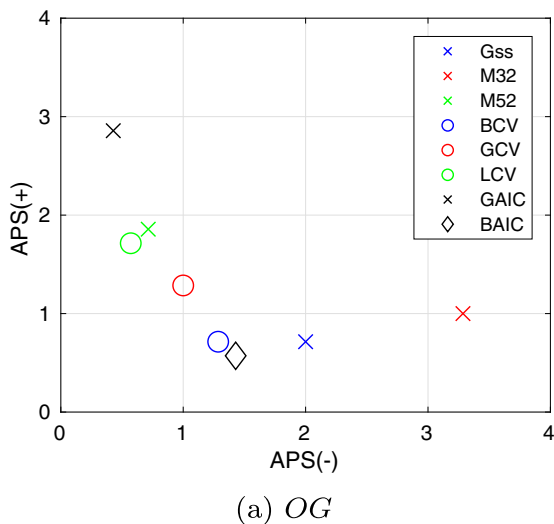
where  $OG_i$  is the optimality gap at iteration  $i$ . Here, a low value of AUC indicates a faster convergence of the optimization algorithm. We apply AUC directly to the  $OG$  instead of its log version so as to better capture the convergence speed in reaching the basin of the global optimum. Statistical tests were performed using Mann-Whitney U-test for the  $\log_{10}(OG)$  and AUC; however, we directly use  $OG$  for non-algebraic problems since we do not exactly know the true optimum for these problems.

In order to quantify the performance of each method relative to other methods, we use the average performance

score (APS) from Bader and Zitzler (2011) that employs the  $p$ -value and statistical significance for all respective methods. The APS metric is fairly simply computed. With  $S$  methods to be compared on  $Q$  problems, the APS value for method  $i$  can be computed by firstly counting its  $j$ -th performance score (i.e.,  $PS_i^j$ ), that is, the number of methods that significantly outperform method  $i$  for problem  $j$ . This procedure is then repeated for all  $Q$  problems to obtain  $PS_i^1, \dots, PS_i^Q$ , where the APS for method  $i$  is simply computed by

$$APS_i = \frac{\sum_{j=1}^Q PS_i^j}{Q} \tag{22}$$

Besides the standard APS, we also compute the number of methods that are significantly outperformed by a specific method. We believe that computing the performance from this viewpoint is necessary since the standard APS measures only one side of the coin. We call this metric APS(+), where



**Fig. 1** Plots of APS(-) versus APS(+) for noiseless synthetic problems



a higher value indicates a better method, and the original APS as APS(-).

For the following explanation of the result, we denote the global ensemble with CV, AIC and the CV-based local ensemble as GCV, GAIC, and LCV respectively. The model selection that selects the lowest CV and AIC are denoted as BCV and BAIC, respectively. Furthermore, EGOs with single kernel function are denoted as Gss, M32, and M52 for the Gaussian, Matérn-3/2, and Matérn-5/2, respectively. Since APS(-) and APS(+) might conflicting with each other, we perform the analysis by plotting the former versus the latter in a multi-objective optimization-like manner. That is, the best methods are located in the Pareto front of APS(-) and APS(+) (notice that we wish to minimize and maximize

the former and the latter, respectively). Here, the APS values were computed separately for noiseless and noisy problems since it is more informative if we analyze these clusters of problem separately.

### 4.2 Results and discussion for synthetic problems

#### 4.2.1 Results for noise-free synthetic problems

The APS(-) and APS(+) of *OG* and AUC for the noiseless problems are shown in Fig. 1a and b, respectively. The boxplots of *OG* and AUC for the noiseless synthetic problem are shown in Figs. 2 and 3, respectively. For each boxplot, we show the performance score at the left side of

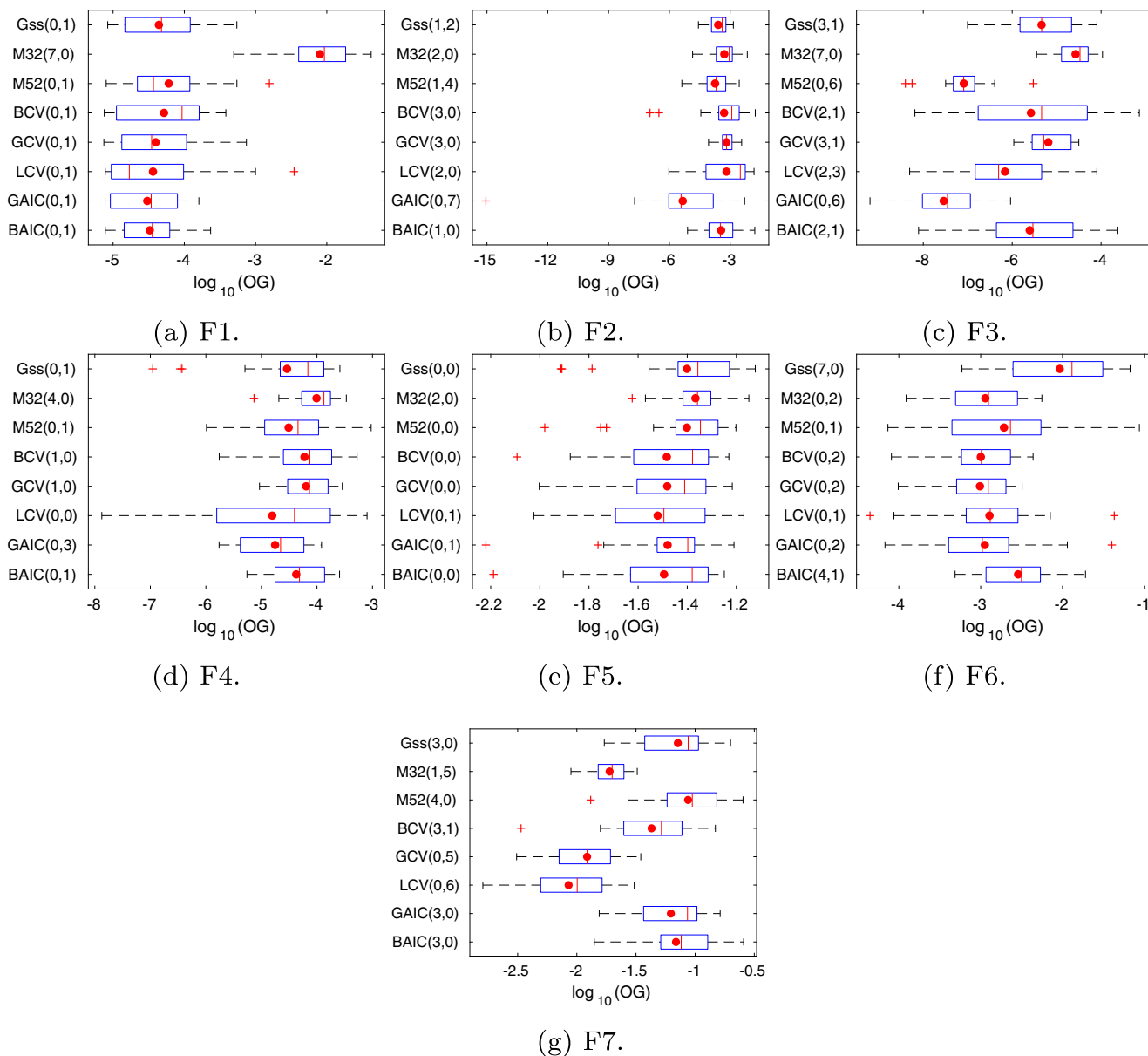
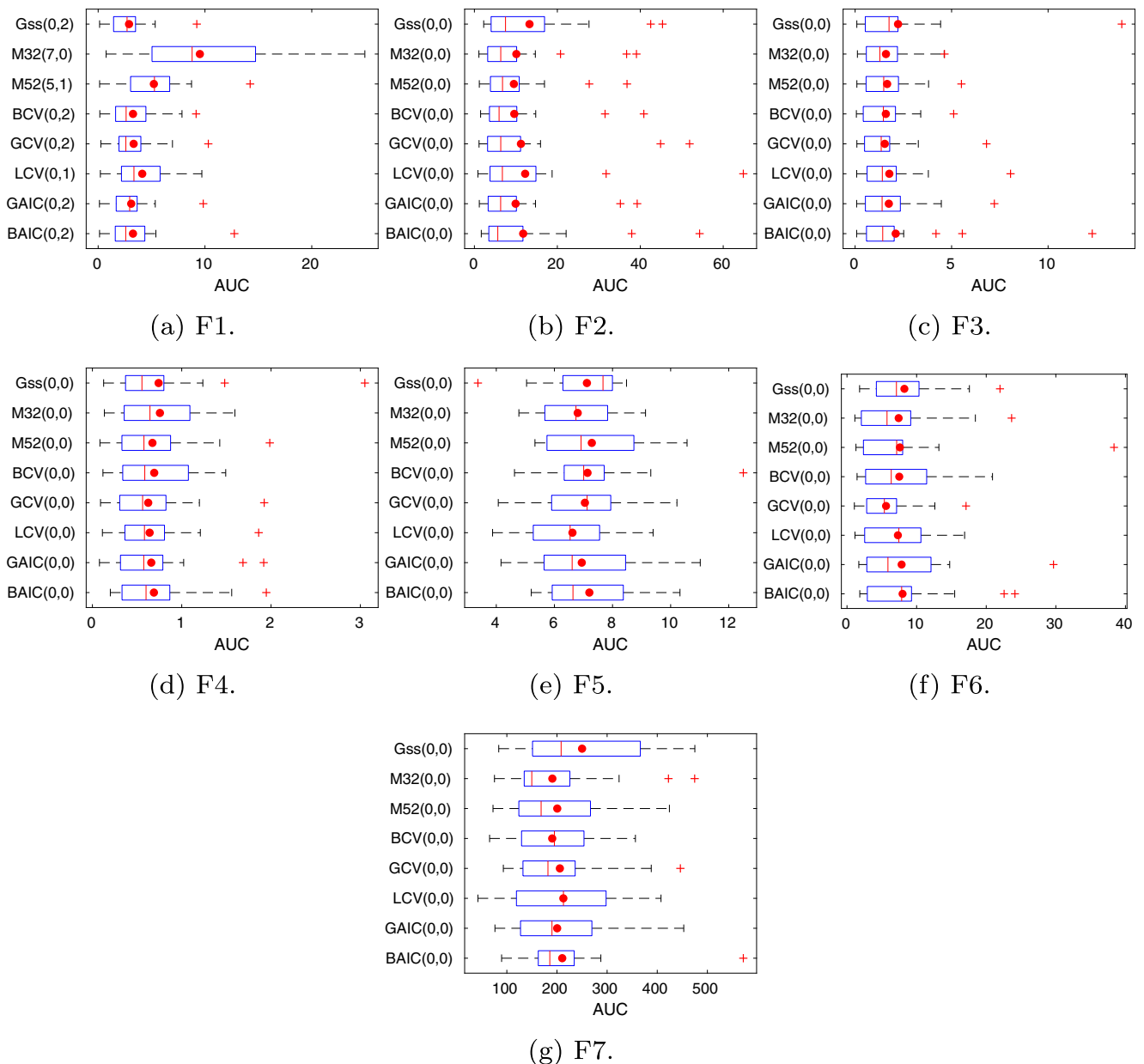


Fig. 2 Boxplots of optimality gap for noiseless synthetic problems



**Fig. 3** Boxplots of AUC for noiseless synthetic problems

the boxplot in the format of *Name of method* (PS(-),PS(+)). For example, BCV(0,1) is for BCV method with the PS(-) and PS(+) value of 0 and 1, respectively.

By comparing the performance of EGO variants with single kernel function from the viewpoint of the optimality gap, it is revealed that EGO with Matérn-5/2 kernel is better in terms of both APS(-) and APS(+) than that of Matérn-3/2 and Gaussian kernel on the noiseless problems considered in this paper. EGO with Gaussian kernel generally outperforms that with Matérn-3/2 kernel except on F6 and F7. The existence of the steep ridge near the global optimum of F6 is obviously the challenge for EGO-Gaussian due to its smoothness assumption. Furthermore, the relatively poor

performance of EGO with the Gaussian kernel on F7 is also surprising. When considering the convergence speed by observing the AUC results, all methods perform similarly except for F1; here, EGO-Gaussian shows significantly faster convergence speed compared to the other single kernel EGOs. Our analysis shows that Matérn-3/2 kernel is not suitable for approximating F1 and yields poor approximation performance which, in turn, decreases its convergence speed. In light of the results, we infer that EGO with Matérn-5/2 kernel is the most robust single-kernel method for solving noiseless problems. We also observe that, in general, the choice of kernel yields small impact on the convergence speed for the noiseless problem except

in cases where different kernel significantly impacts the approximation power.

Comparing the results of EGOs with multiple kernels from the viewpoint of optimality gap, the average performance of GCV is outperformed by that of LCV. The most plausible explanation for this is that GCV does not possess good control in capturing the locality of the objective function, which is an important requirement for optimization; it is also worth noting that all synthetic problems except the sphere function have nonlinear responses to various degrees. The fact that LCV provides good solutions in all noiseless synthetic problem indicates that it has a more precise accuracy control near the global optimum, which explains why it outperforms GCV. GCV is particularly poor for low-dimensional problems (F1-F4) but its performance is relatively good on medium and high-dimensional problems (F5-F7); it is capable of avoiding kernel misspecification on F5 and F6 and even outperforms all EGOs with a single kernel on F7. The results also suggest that the dimensionality of the problem influence the performance of GCV.

The best performer for the noiseless problem is GAIC; it yields significantly higher performance on F2, F3, and F4, outperforming EGOs with single kernel function. GAIC also performs strongly on F1, F5, and F6. GAIC even strictly outperforms EGOs with single kernel functions on F3 and F4 functions which means that mixing multiple Kriging models could produce significant advantages over solely using a Kriging with a single kernel. The only instance where the performance of GAIC is not satisfactory is on the F7 function; regardless, its performance is still better than the corresponding worst performer (i.e., EGO with Matérn-5/2 kernel).

We observe that there is no significant advantage obtained from utilizing BCV and BAIC. In this respect, both BCV and BAIC are dominated by all ensemble approaches as one can observe in the APS(-) vs APS(+) plot for the optimality gap. Indeed, there is a risk of choosing an inappropriate kernel even when model selection is utilized.

In some functions, the application of model selection yields performance worse to that of EGOs with a single kernel (for example, the result of BCV and BAIC for F2 and F6, respectively). Both BCV and BAIC yield good performance only on functions where the choice of kernel significantly impact the Kriging approximation quality as in F1. However, both BCV and BAIC are unable to outrank the best performing EGO with a single kernel; this means that model selection only further complicates the issue and offering no panacea to the problem of kernel selection. Conversely, global ensemble techniques include both the good and poor kernel in the formulation and this creates the robustness in its performance or even yield a better result.

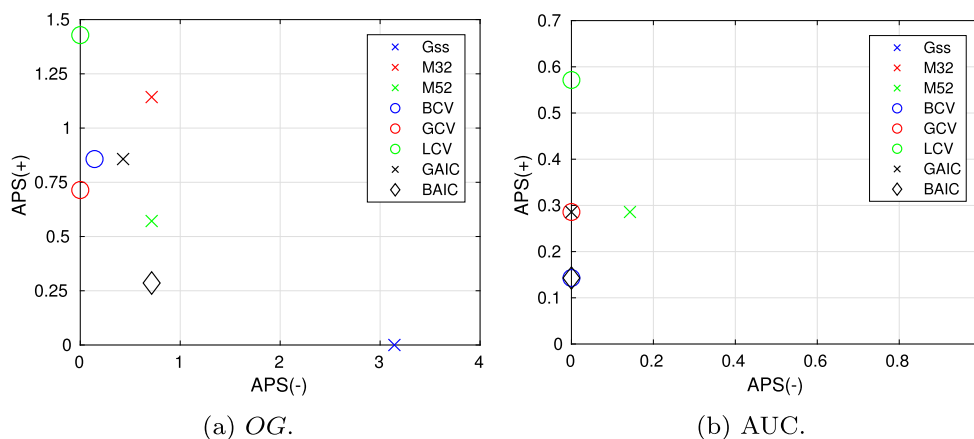
The fact that the performance of BAIC is worse than GAIC is rather interesting. On the one hand, the combination of Kriging with various kernel functions through Akaike weighting leads to a highly performing EGO scheme. On the other hand, it is the opposite case for BAIC. In this regard, the performance of BAIC seems to be limited by the best performing kernel for a given particular problem; however, it turns out that it is difficult to select the best performing kernel via model selection. In light of the results for synthetic problems, we infer that GAIC is a highly robust and high performing method for solving smooth and nonlinear problems. The more desirable performance of GAIC is also accompanied by its simplicity; that is, the operation to compute the weights are fairly simple. Nevertheless, it is necessary to analyze the performance of each method on noisy and non-algebraic problems before drawing general conclusions.

### 4.2.2 Results for noisy synthetic problems

The results for synthetic problems with 1% noise level are shown in Figs. 4, 5, and 6, while the results for the 5% noise level are depicted in Figs. 7, 8, and 9.

When comparing EGOs with a single kernel for handling noisy problems, the first obvious fact that we

**Fig. 4** Plots of APS(-) versus APS(+) for noisy synthetic problems with 1% noise level



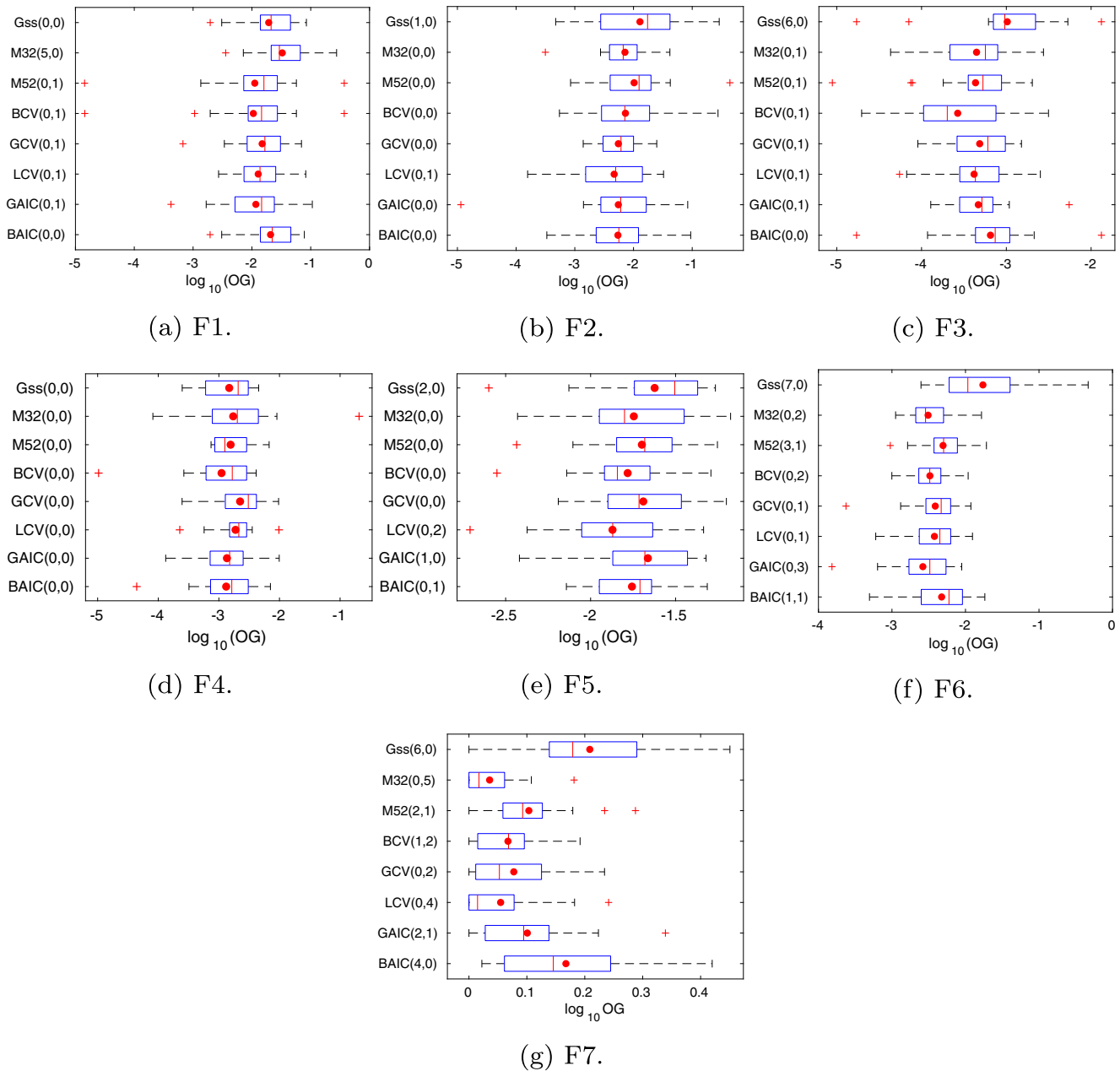
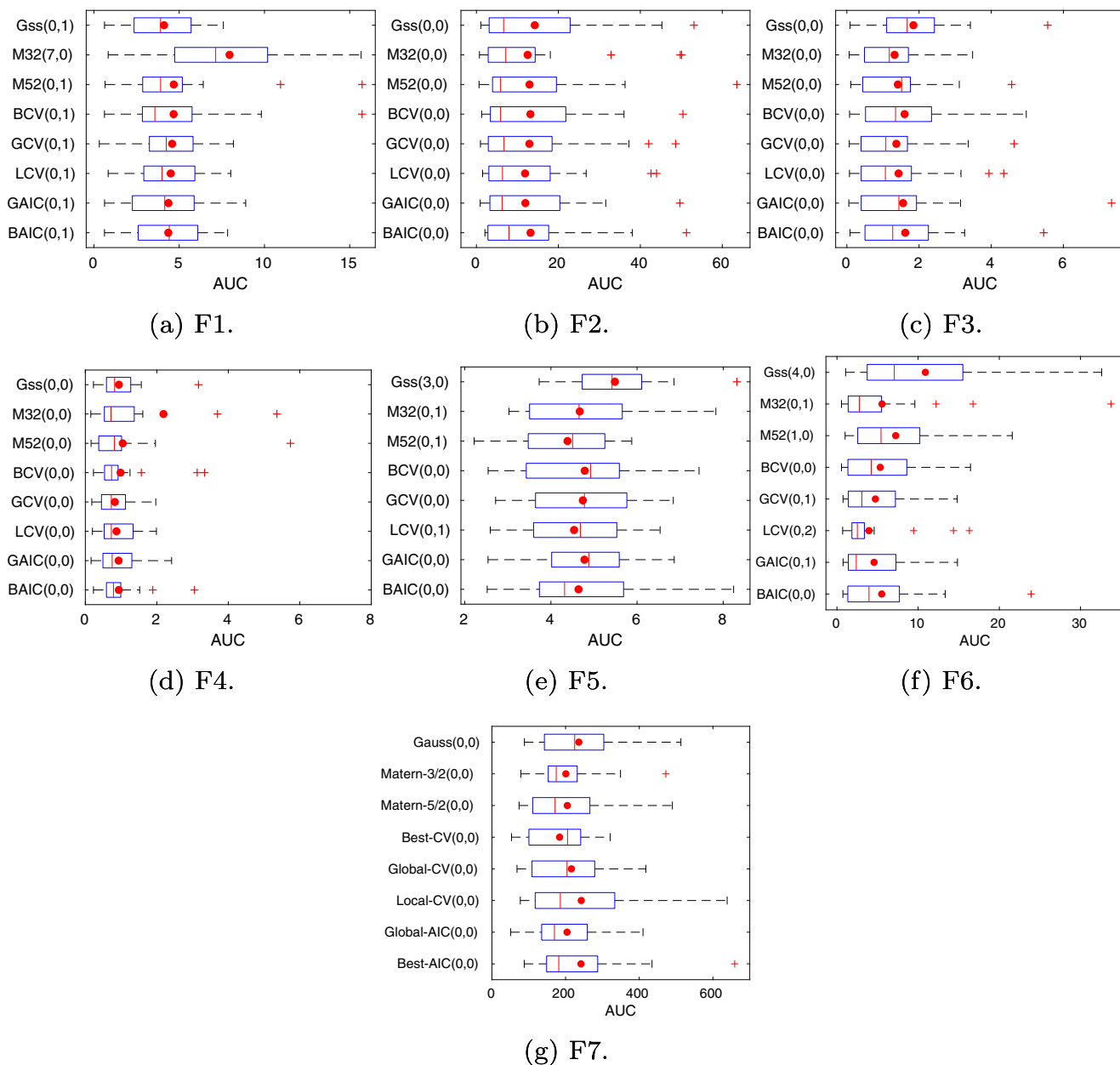


Fig. 5 Boxplots of optimality gap for synthetic problems with 1% noise level

observe is that EGO with Gaussian kernel yields relatively poor performance in terms of both optimality gap and convergence speed. On the other hand, both Matérn kernels yield relatively good performance when applied on EGO, especially the Matérn-3/2 kernel. The only instance where EGO with Gaussian kernel performs relatively better than that of Matérn-3/2 is on the F1 function with 1% noise level; which is due to the characteristic of Branin function that suits the applicability of Gaussian kernel. However, the use of single Gaussian kernel yields no benefit when the noise level is moderate. The smoothness assumption imposed by Gaussian kernel proved to be problematic and this affects

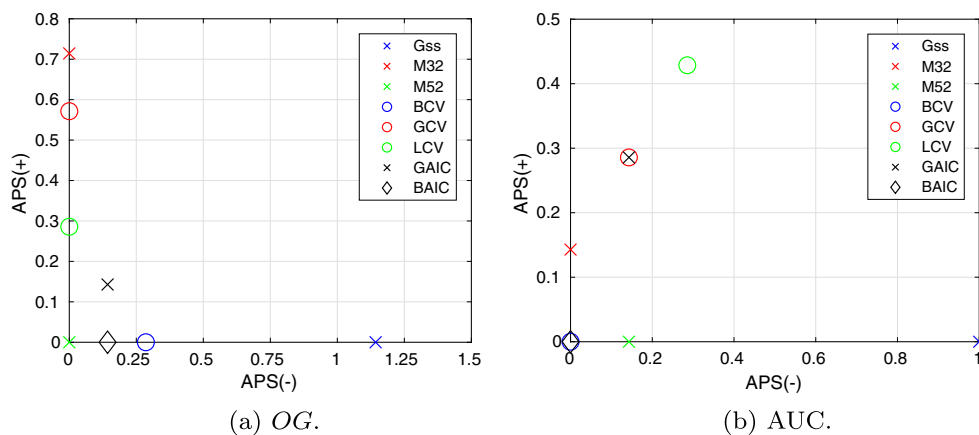
its performance on tackling noisy problems. In light of the results, it is safe to conclude that EGO with single Gaussian kernel is not really suitable for handling noisy problems. Thus, we suggest deploying the Matérn kernels when using single kernel EGO for handling real-world problems that possibly feature noise. Furthermore, such observation is also in line with Stein’s recommendation to use Matérn kernels in the context of general approximation (Stein 2012).

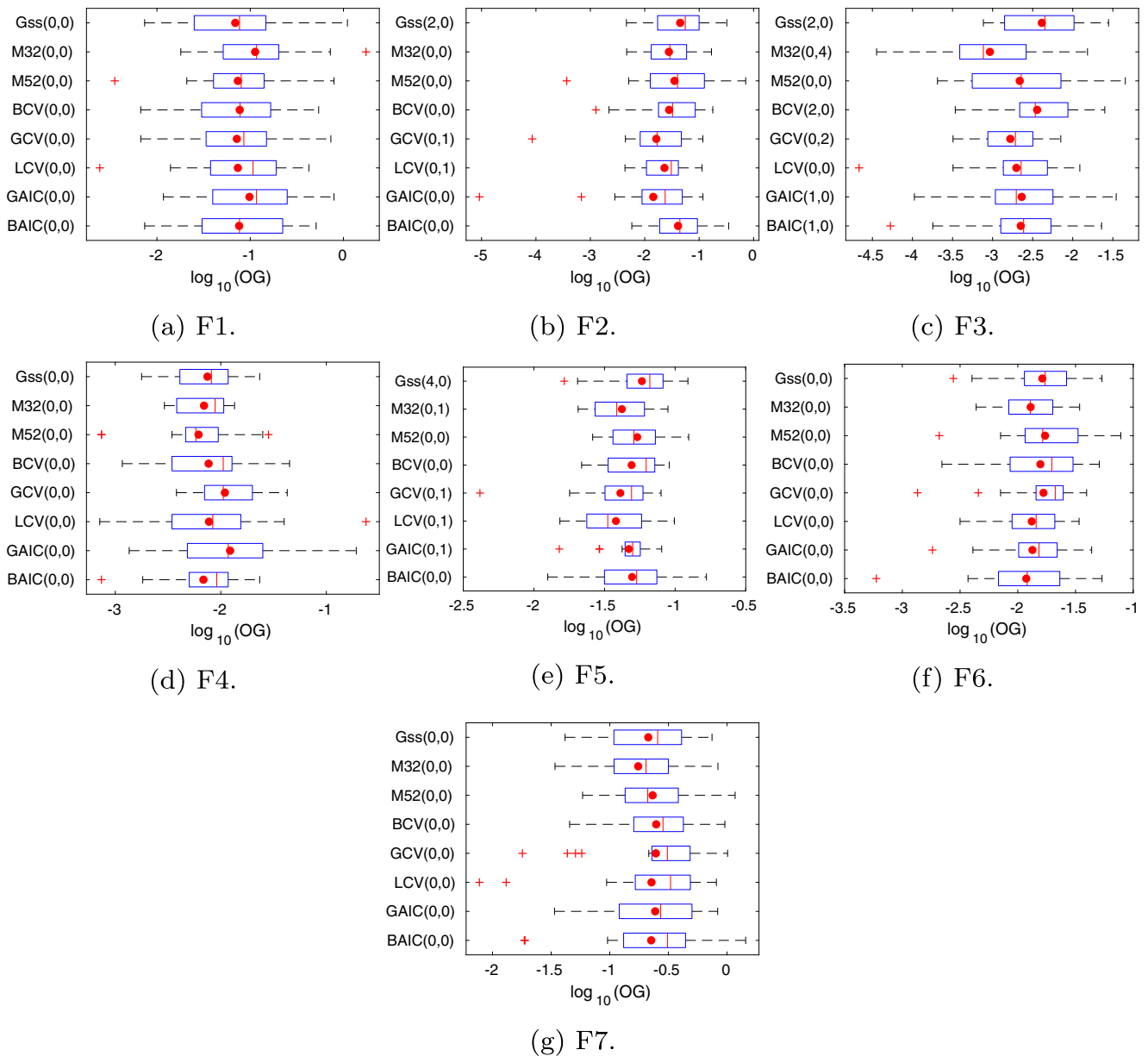
In some noisy problems, there are clear differences in convergence speed between all methods, which is particularly notable on medium and high-dimensional problems. However, still, no such differences in AUC



**Fig. 6** Boxplots of AUC for synthetic problems with 1% noise level

**Fig. 7** Plots of APS(-) versus APS(+) for noisy synthetic problems with 5% noise level



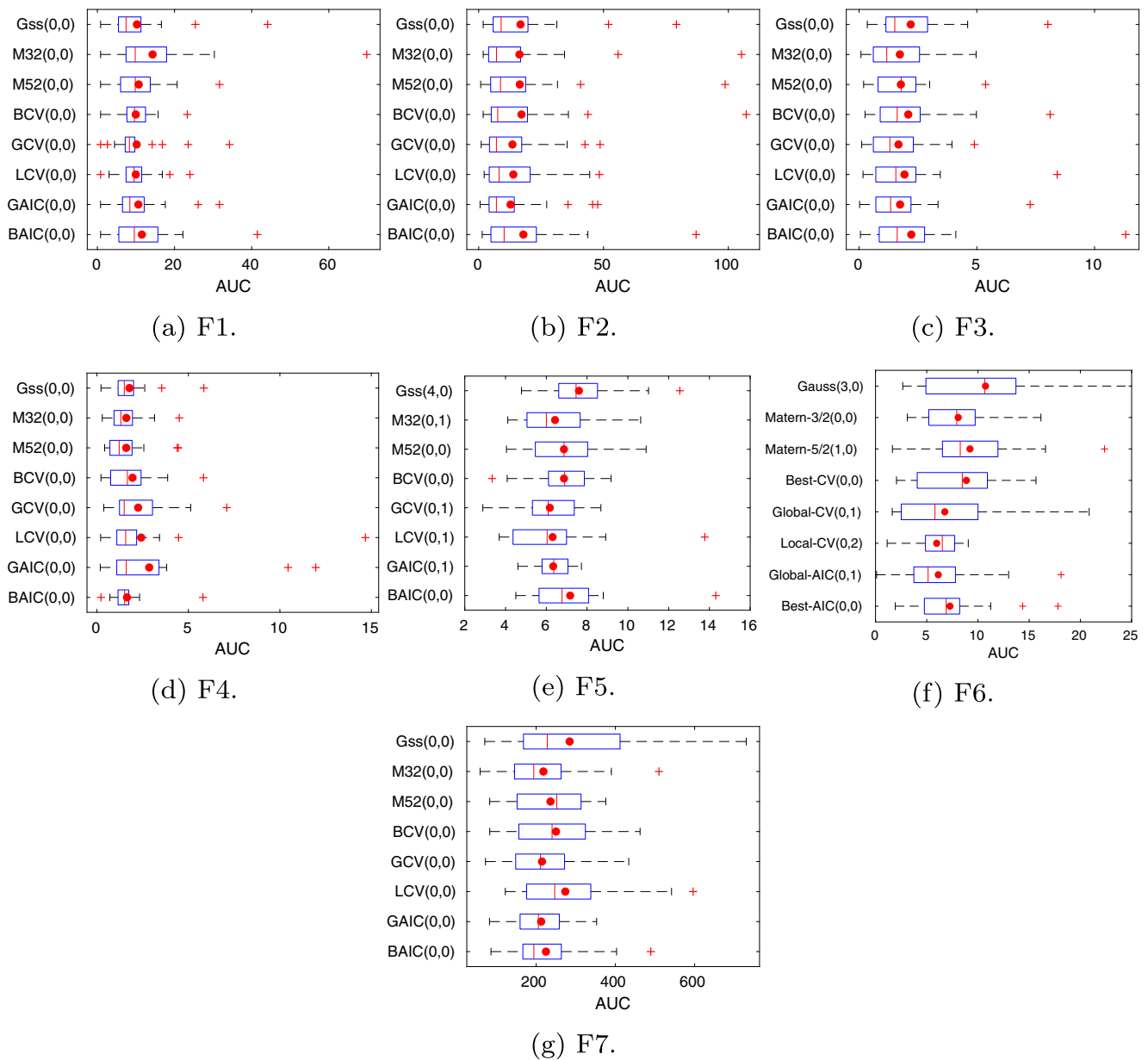


**Fig. 8** Boxplots of optimality gap for noisy synthetic problems with 5% noise level

were observed for F2, F3, and F4, regardless of the noise level. That is, in low-dimensional problems except for F1, we observe that the choice of kernel primarily affects the final convergence (i.e., exploitative power) and not the convergence speed, which is logical since it is easier to detect the basin of the global optimum in low-dimensional problems. In contrast, EGO needs more function evaluations to explore the high-dimensional design space before arriving into the basin of global optimum. That is to say that the presence of noise, albeit slightly, affects the explorative power of EGO with various choices of kernel. Especially, the impact of noise on convergence speed can be clearly observed on F6

where there are clear differences between the performance scores.

From the viewpoint of model selection and ensemble of surrogates, the most obvious difference between the results for noisy versus those of synthetic problems is that GAIC loses the powerful trait that makes it the strongest performer on smooth synthetic problems. In this sense, the convergence speed of GAIC is not significantly affected but its optimality gap does. It can be seen that the APS(-) and APS(+) of GAIC in terms of optimality gap becomes worse relative to the other techniques as the noise level increases. This indicates that the exploitative capability of GAIC diminishes as the noise level increases. The results



**Fig. 9** Boxplots of AUC for synthetic problems with 5% noise level

suggest that AIC is not a really useful measure as a selection or mixing criterion when noise is present. Conversely, methods that are based on CV error (i.e., BCV, GCV, and LCV) surprisingly yield better performance on noisy problems compared to that of smooth problems. For low levels of noise, LCV appears as the best performer that generally dominates over the other methods from both the optimality gap and convergence speed viewpoint; although the difference in the latter is not so significant. The fact that LCV performs well in problems with low noise level indicates that the local measure used to compute the non-constant weight function in LCV is still useful. However, such local measure is not really helpful when the noise level

is moderate due to the difficulty in distinguishing between the true response and the noise. It turns out that the global average of CV error is more useful for moderate noise level as indicated by the higher performance of GCV as compared to LCV on such problems.

The better performance of GCV on moderate noise level indicates that it is more sensible to trust the CV error information rather than AIC while constructing the ensemble of Kriging models in the context of EGO. The fact that BAIC yields performance that is worse than that of GAIC further suggests that BAIC tends to select improper kernels during the model selection process; however, it is still a better choice than the Gaussian kernel. The results

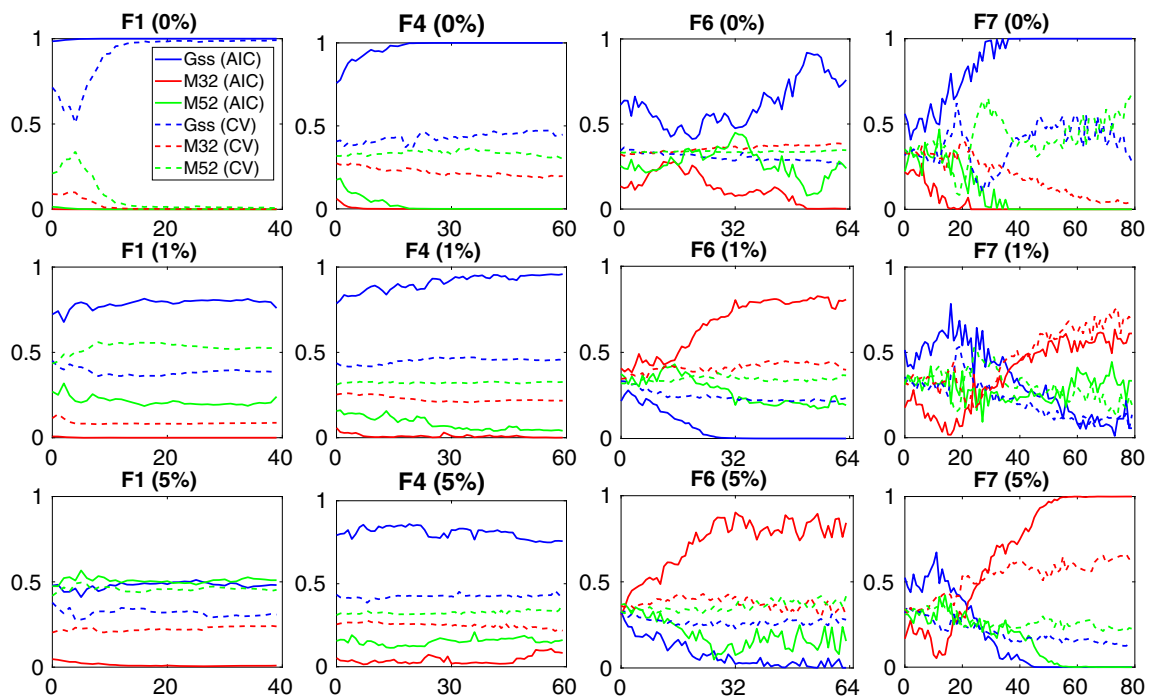
demonstrate that model selection based on AIC should be avoided for solving problems in a noisy environment. It is also worth noting that it is difficult to predict the level of noise, particularly in high-dimensional problems. In this regard, combining Kriging models with AIC is a better remedy than BAIC since well-performing kernels are included in the process; although it is worth noting that GAIC yields performance worse than that of GCV in a moderate level of noise.

It is also clear to see that performance differences between all techniques become less for a moderate level of noise; one can see a smaller range of APS(-) and APS(+) in the optimality gap, as compared to the problems with a lower noise level. It is then safe to assume that differences in performance would further diminish as the noise levels increase. The reason for this is that it would become more difficult for the Kriging model to distinguish the true response and the noise, which particularly affects how the Kriging model and EGO perceive the location of the true optimum. Nevertheless, some clear differences still exist; most notably the relatively poor performance of the Gaussian kernel. Although the Matérn-3/2 kernel yields the best performance on the moderate noise level, note that it outperforms the Matérn-5/2 kernel in terms of optimality gap only on F3 and F5. In the previous research, Picheny et al. (2013) observe that the impact of the choice of the kernel is not significant while our experiments indicate the opposite. However, it is worth noting that their experiments

were performed in a highly noisy environment (i.e., 5%, 20% and 50% noise level) where it is likely that the choice of the kernel does not really affect the EGO performance. Conversely, our experiment also considers low noise levels (i.e., 1%) which shows that the choice of kernel indeed affects the performance. Moreover, in our experiment, such differences also can be observed for the 5% noise level.

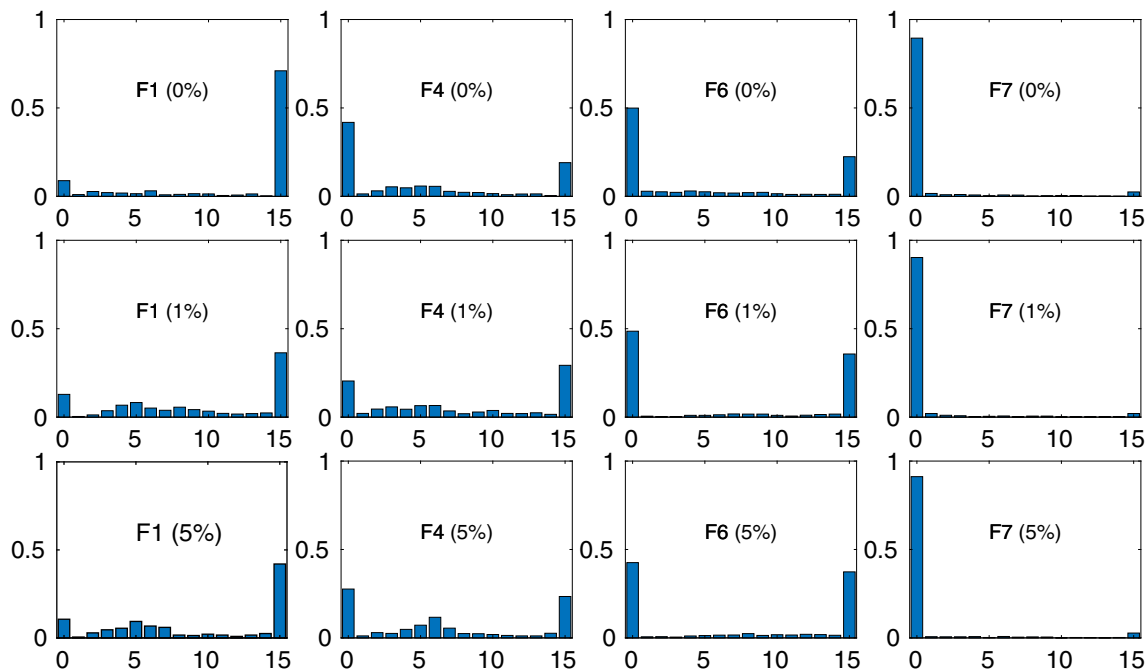
#### 4.2.3 Synthetic problems: weight analysis

We analyze the weight generated by GAIC and GCV in order to gain important insight regarding their mechanism in seeking the optimal solution. The differences in performance between GCV and GAIC, despite both utilizing a constant weight methodology, can only be attributed to how they determine the weight at each iteration of EGO. To this end, we studied the weights generated during the optimization process as shown in Fig. 10, which shows the mean values of the weight, iteration-wise. For LCV, we do not depict the generated weights since they are highly variable over the design space. Instead, we depict the attenuation coefficients generated by LCV in order to see the influence of locality (see Fig. 11). We limit the discussion only on the global and local ensemble techniques since the performance of BAIC and BCV are generally lower than the ensemble techniques when averaged across all problems and noise levels. In our analyses, the weight and attenuation coefficient plots for F2 and F3 are not



**Fig. 10** Weight history of the global ensemble approaches for F1, F4, F6, and F7 with various noise levels as written inside the brackets. Here,  $x$ -axes and  $y$ -axes denote the number of EGO iterations and the weights, respectively





**Fig. 11** Histogram of attenuation coefficients called by LCV on synthetic problems with various noise levels as written inside the brackets. Here,  $x$ -axes and  $y$ -axes denote the attenuation coefficients and the normalized frequencies, respectively

shown, since their characteristics are similar to those of F1. Similarly, the plots for F5 are also not shown due to their similarity to those of F6.

For noiseless problems, we observe that GAIC tends to converge toward a single kernel function on the long run. On all noiseless problems, Gaussian kernel clearly dominates the composition in GAIC especially at the middle and at the end of the search. One fact worth noting is that there is a process of kernel mixing before the weights converged to a Gaussian kernel; this trend is obviously strong for F7. It is interesting to see that GAIC outperforms EGO with single kernel despite that the former tends to select Gaussian kernel in a long run. Except for F1, i.e., the function where EGO with a Gaussian kernel strictly dominates those with Matérn kernels, GCV tends to generate mixing weights that differ in an extreme way compared to GAIC. To be exact, GCV tends to be conservative in determining the weight for the global ensemble. In light of these results, we infer that the initial mixing of Kriging models at early iterations produces a favorable effect on the exploration phase for solving noiseless problems; this, in turn, also leads to a highly efficient exploitation phase. This testifies that, at least in noiseless synthetic problems, GAIC has a better mechanism to find the ‘correct’ weights, compared to its CV counterpart.

The introduction of noise notably affects the way GCV and GAIC generate the mixture proportion. In general, the presence of noise reduces the proportion of Gaussian kernel

and strengthens that of the Matérn kernels in the mixture, especially for GAIC. This trend can be clearly observed in the weight results for the F1 and F7 functions; for F7, the Matérn-3/2 fully replaces the Gaussian kernel on the 5% noise level case. This indicates, as expected, the presence of noise decreases the appropriateness of Gaussian kernel and the scheme tends to incline more toward the Matérn-kernel, which is particularly designed for handling noisy problems. Unfortunately, the performance of GAIC decreases as the noise level escalates. On the other hand, similar to the results for noisy synthetic problems, GCV stays conservative in determining the mixture proportion. The only observable significant changes in the weight proportion of GCV is on the F1 function, which is primarily caused by the decrease in approximation quality of Kriging with the Gaussian kernel, as the noise increases. Nevertheless, the conservativeness of GCV tends to be advantageous for noisy problems since it yields good performance on both noise levels.

As can be observed from Fig. 11. The pointwise ensemble yields an observed strong effect on reducing the CV error and produces fundamentally different surrogate models compared to those of the pure global ensemble. Furthermore, we also observe that the attenuation coefficients tend to concentrate on the extremes (i.e.,  $\Theta = 0$  or  $\Theta = 15$ ); which indicates that the models tends to either mimic the global ensemble or produce local ensembles with a strong locality effect. However, the proportion of LCV Kriging

models with intermediate values of  $\theta$  is also substantial; it appears not so significant since they are fairly distributed between  $\theta = 1$  and  $\theta = 14$ .

The locality effect is particularly strong in low dimensional problems which explain why LCV outperforms GCV in noiseless synthetic problems. We also observe that LCV tends to mimic the global ensemble as the dimensionality increases. This is because it becomes more difficult to predict the appropriate non-constant weight function due to the sparsity of sampling points in the high-dimensional design space (i.e., the curse-of-dimensionality). However, the pointwise ensemble still takes notable effect as can be observed from the appearance of  $\theta$  that is higher than zero even for F7 problem. In fact, this small proportion of LCV-Kriging models aids in local exploitation of the F7 function that made it outperforms GCV on noise level of 0% and 1%. Furthermore, results show that no clear trend regarding the impact of the noise level on the attenuation coefficients of LCV can be observed; thus, no solid conclusion could be drawn regarding this issue. However, it is possible that the presence of noise further strengthens the effect of locality. At least for F4 and F6, the proportion of LCV model with  $\theta = 0$  decreases when noise is introduced to the synthetic problems. This is reasonable since noise creates kinks on the response surface which

makes it appears that the problem features stronger local activity.

### 4.3 Results and weight analyses for aerodynamic optimization and support vector regression problems

Results for non-algebraic problems are shown in Figs. 12 and 13. Note again that we investigate both the optimality gap and convergence rate since it is important to perform the analysis from both viewpoints. For the non-algebraic problems, we do not plot the APS(-) and APS(+) since it is better to perform individual analysis for each non-algebraic problem rather than computing the average of them.

Results for non-algebraic problems show that, in general, EGO with the Matérn-3/2 kernel yields better performance compared to that of Gaussian and Matérn-5/2. In particular, EGO with Matérn-3/2 kernel is significantly better than the other single kernel-based EGOs on case A2 and A3. However, the EGOs with Matérn-5/2 and Gaussian kernel are better than that of the Matern-3/2 on A5, which indicates that there is no single best performing kernel. One can also see that there are notable outliers in case A2 in terms of optimality gap, which indicates that the global optimum is difficult to find and is probably located in a narrow valley

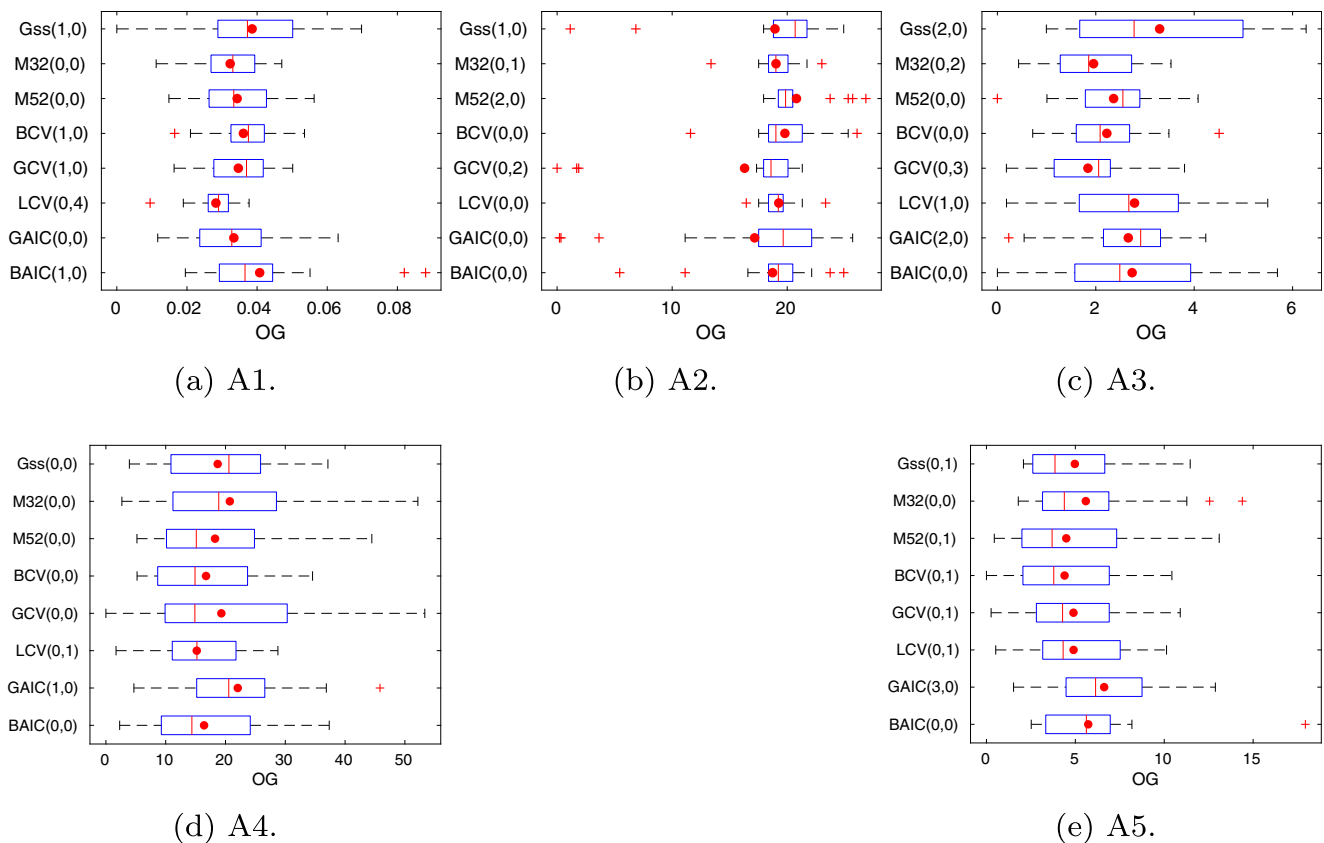


Fig. 12 Boxplots of optimality gap for non-algebraic problems

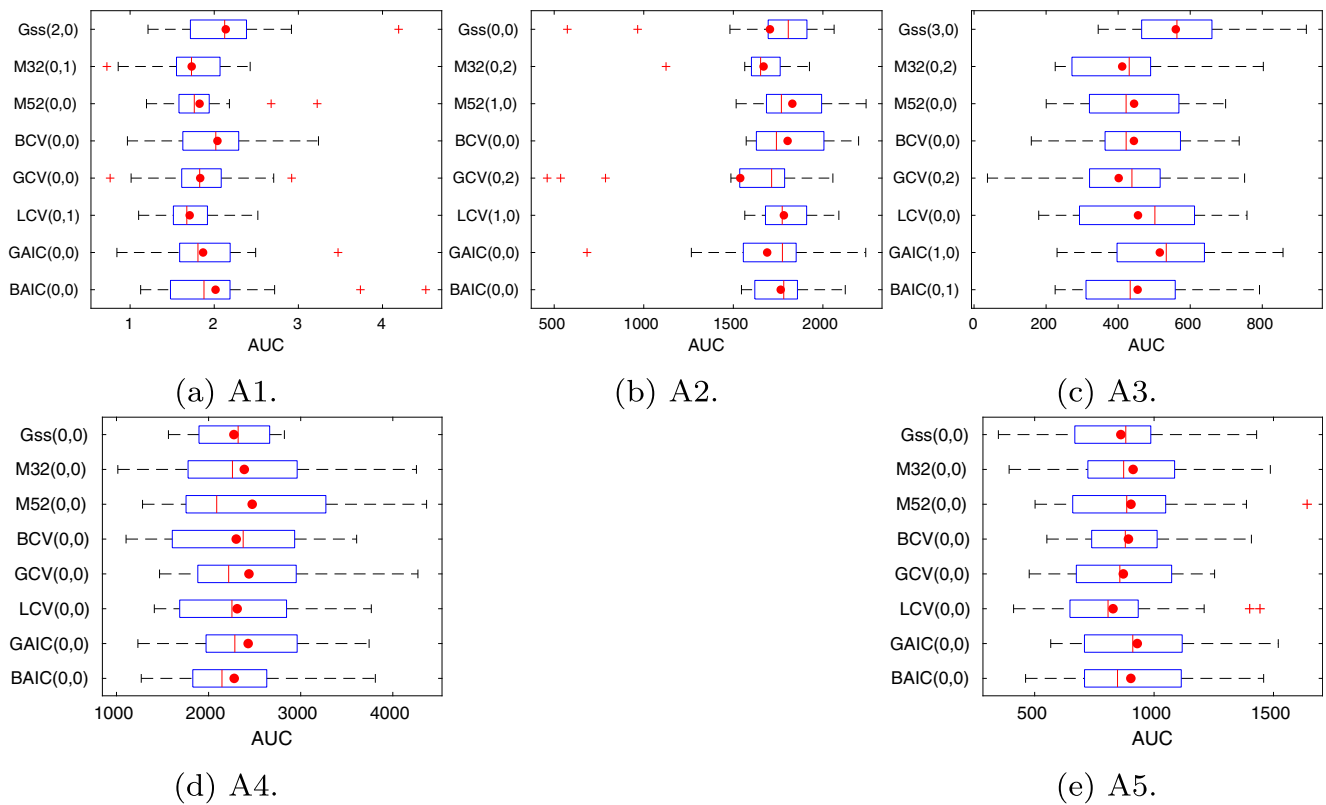


Fig. 13 Boxplots of AUC for non-algebraic problems

similar to that of the Rosenbrock function. The use of single Gaussian kernel in EGO is not recommended as evidenced by the worst performance in three out of five non-algebraic problems (i.e., A1, A2, and A3). The convergence speed is also affected by the choice of kernel as can be seen in case A1, A2, and A3; again, EGO with Matérn-3/2 kernel yields the best convergence speed on the aforementioned problems.

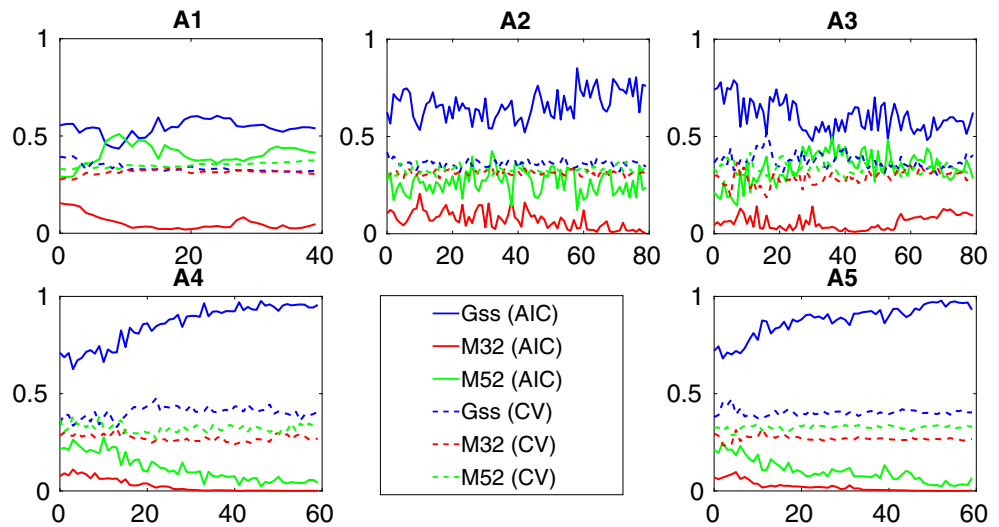
Two ensemble approaches, i.e., LCV and GCV, take the spotlight for the non-algebraic problems. On the one hand, LCV bested the other methods on A1 with regard to the optimality gap and convergence speed; LCV also outperforms the others in terms of the optimality gap for case A4. On the other hand, GCV yields the best performance on A2 and A3 as evidenced by its lowest APS(-) and highest APS(+) for the two performance metrics. GAIC and BAIC display performances that are not better than those of EGO with the Gaussian kernel. One possible reason for the poor performance of GAIC is that it puts too much weight on Gaussian kernel, in which the Gaussian kernel proved to be an unsuitable kernel for the non-algebraic problems considered in this paper. We also observe that there is no gain attained from utilizing BAIC and BCV for non-algebraic problems.

The weight history plots depicted in Fig. 14 show an interesting trend. For A4 and A5, we observe a converging

trend of the mixture proportion towards one kernel (i.e., Gaussian) for GAIC similar to that of synthetic problems; however, such trend is not observed in A1, A2, and A3. Such non-converging trends indicate that the disparity in AIC for the Kriging models with different kernels do not change much as the EGO progresses. Regardless, GAIC does not yield satisfactory solutions on all non-algebraic problems. On the other hand, GCV is still a conservative technique in terms of determining the weight for all non-algebraic problems. However, the conservativeness of GCV tends to be a positive feature since it yields good performance, especially on problems A2 and A3. The fact that GCV outperforms the best performing kernel on case A2 and A3 signifies that there are extra benefits obtained by utilizing the global ensemble of Kriging models with different kernels in non-algebraic problems, which can only be attributed to the improvement in the quality of the surrogate model.

The histogram of attenuation coefficients for the non-algebraic problems are shown in Fig. 15. Here, a trend similar to that of synthetic problems is observed; that is, the local influence of design points is stronger on the low-dimensional problem (i.e., A1) compared to high-dimensional problems (i.e., A2, A3, A4, and A5). The strong performance of LCV on case A1 can then be attributed to the capability of LCV in capturing the local activity in this particular case.

**Fig. 14** Weight history of the global ensemble approaches for non-algebraic design problems. Here,  $x$ -axes and  $y$ -axes denote the number of EGO iterations and the weights, respectively



**4.4 Discussion and remarks**

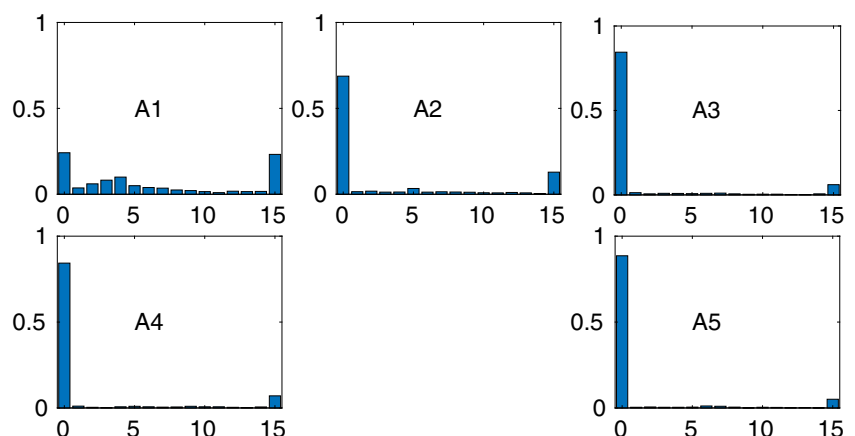
We highlight some important points in light of the results. Here, our aim is to summarize the overall findings to serve as a guideline and also derive some recommendations for a more efficient application of the mixing or model selection of Kriging with multiple kernel functions. Moreover, we also synthesize the finding from experiments on synthetic and non-algebraic problems.

1. **There is no single best performing kernel.** That is, the best kernel depends on the characteristics of the problem and the noise level; however, we do have recommendations. For the noiseless problems considered in this paper, Matérn-5/2 is the best performing kernel, that surpassing Gaussian and Matérn-3/2. However, when noise corrupts the black-box function, Matérn-3/2 outranks the others regardless of the noise level. Even EGO with Matérn-3/2 generally outperforms the ensemble and model selection techniques for moderate

noise levels. On the other hand, in light of the results, we do not recommend using Gaussian kernels, since it yields results that are typically worse compared to the other kernels on synthetic and non-algebraic problems. It is worth noting that when we say worse it does not necessarily mean that its performance is poor, it only means that there are other methods that surpass its performance. In light of these results, we recommend the use of Matérn kernels instead of Gaussian for handling real-world problems when one wishes to deploy a single kernel in EGO.

2. **Model/kernel selection offers no advantage compared to the single kernel EGO.** Using model selection is perilous in the sense that it is likely that the selection scheme chooses non-optimal kernels during the optimization process as evidenced by the generally worse performance of model selection compared to the single best-performing kernel. We, therefore, do not suggest the crude use of CV and AIC for online model selection.

**Fig. 15** Histogram of attenuation coefficients called by LCV on non-algebraic problems. Here,  $x$ -axes and  $y$ -axes denote the attenuation coefficients and the normalized frequencies, respectively



3. **In general, ensemble of Kriging methods yields better performance than that of model selections.** Furthermore, ensemble techniques are able to increase the optimization performance in terms of optimality gap and convergence speed, or at least avoiding poor kernels. Nevertheless, the answer to the question of which ensemble technique should be applied depends on the noise level of the problem. In this regard, the global ensemble based on Akaike weight is useful for noiseless problems; however, its performance significantly deteriorates as the noise level increases. On the other hand, CV-based ensemble techniques are more robust to the impact of noise and show satisfactory performance on noisy synthetic and non-algebraic problems. Hence, unless one could really be sure that the problem to be tackled is free of noise, one should deploy either CV-based global or local ensemble techniques for handling real-world problems. Despite its (slightly) more complex implementation as compared to the global ensemble, EGO with CV-based local ensemble (i.e., LCV) performs favorably in all test problems. We particularly recommend to use LCV when the dimensionality is low (although it also shows relatively good performance on medium to high-dimensional problems), not to mention that LCV performs well on both noiseless and noisy problems. On the other hand, GCV is recommended for medium to high-dimensional noisy problems.

As a side note, it is worth noting that not all characteristics of real-world problems can be fully replicated by synthetic problems; hence, whenever possible, tests on optimization of real-world processes/non-algebraic problems should also be performed with proper statistical tests. That is to say that we could learn important insights regarding the capability of optimization methods from tests on synthetic problems, which is beneficial and necessary since they are cheap to evaluate; nevertheless, tests on real-world problems should not be just treated as a mere application but also for gaining important understanding.

## 5 Conclusions

When performing Kriging-based efficient global optimization (EGO), one typically fixes the choice of kernel function in order to approximate the response surface. It is likely that the performance of EGO can be improved through a proper choice or mixing of Kriging models with different kernel functions. In the present article, Kriging with automatic selection and mixture of kernel functions were investigated for solving global optimization problems under limited budget. We investigated model selection based on

CV and AIC in order to identify the most potential kernel function on a set of synthetic and non-algebraic problems. Besides automatic kernel selection, several methods to mix multiple Kriging models were also studied, that is, global, local (i.e., pointwise) ensemble based on cross validation, and global ensemble based on AIC. Three widely used kernel functions (i.e., Gaussian, Matérn-3/2, and Matérn-5/2) are employed to construct the individual Kriging models.

The performance of Kriging models with single kernel function, automatic kernel selection, and the mixture of kernel function were assessed on seven noiseless and noisy synthetic problems and five non-algebraic problems (i.e., aerodynamic design and parameter tuning in SVR). Considering our main objective that attempts to tackle engineering design optimization, the results suggest that there is an evident merit obtained from utilizing the ensemble of Kriging with multiple kernel functions. The most apparent advantage is that the use of the ensemble approach avoids a performance penalty that could be incurred by kernel misspecification. We also observe that a proper use of ensemble technique yields substantial improvements in solution quality and convergence rate as compared to EGOs with single kernel. However, one needs to take into account the impact of noise when choosing the suitable ensemble technique to be employed.

EGO with global ensemble based on AIC bested the others on noiseless synthetic problem. However, its use is not recommended for noisy problems, which might be typical for a majority of engineering design problems. EGO with AIC-based global ensemble is then more suitable for solving nonlinear problems with smooth characteristics, which could be encountered in problems such as the tuning of a neural network or also engineering problems in situations where one could really suppress numerical noise. On the other hand, CV-based ensemble techniques, global or local, yield considerable performance and robustness in noisy synthetic and non-algebraic problems. We also note that the EGO with local ensemble is a highly performing method regardless of the noise level and dimensionality; in this regard, the CV-based local ensemble possesses more control on the local surface than its CV-based global ensemble counterpart.

In the future, we hope to perform tests on more instances of real-world processes from various engineering and research disciplines. Such tests would further improve our understanding of the capabilities, and thus also the pitfalls, of the ensemble of Kriging with multiple kernels on solving real-world problems. We acknowledge that our work is still limited in terms of parallelization in the sense of batches of function evaluations, although parallelization still could be done for the function evaluation itself. Future work will include this type of parallelization,

especially regarding the most effective technique to perform such parallelization. Another interesting research avenue is to consider the mixture of kernel directly inside the Kriging model and compare it with the present approaches. Finally, further development for handling multi-objective optimization problem is to be considered in the near future.

**Acknowledgments** Koji Shimoyama was supported in part by the Grant-in-Aid for Scientific Research (B) No. H1503600 administered by the Japan Society for the Promotion of Science (JSPS).

### Appendix A: Non-algebraic optimization test problems

This appendix details the non-algebraic cases used as benchmark problems for this paper. The objective function for the four aerodynamic cases is to maximize  $L/D$ . Note that the flight conditions for all aerodynamic cases is explained in Section 4.1.1.

#### A.1 Case A1: support vector regression parameters tuning test problems

This problem is the parameter optimization of a SVR model in order to minimize the CV error; that is, the task is to construct an SVR model that generalizes the data set well. Here, we use the abalone data set that contains the physical measurements of abalones with 4177 samples, 1 categorical predictor, 7 continuous predictors, and an integer response variable. Although this problem is not an engineering problem, it serves as a good test problem due to its nonlinear response surface characteristic and moderate level of noise; which we believe represent some engineering problems in terms of the input-output relationship.

The decision variables are the box constraint parameter and kernel scale which are defined in the  $[10^{-3}, 10^3]^2$  space in a logarithmic scale. For this problem, the objective function is to minimize the 5-fold CV error of the SVR model. The problem is relatively fast to evaluate due to the relatively low size of the training set used to construct the SVR model. Furthermore, the problem exhibits a relatively low level of noise which means that Kriging

**Table 2** The upper and lower bounds for subsonic airfoil optimization problems (case A2 and A3)

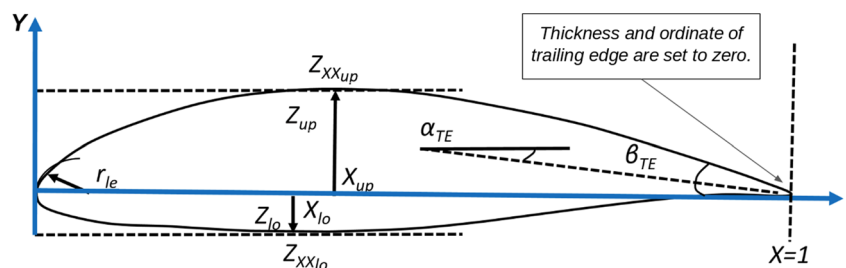
Variable	Definition	$l_b$	$u_b$
$r_{le}$	leading edge radius	0.0108	0.0162
$X_{up}$	upper crest position in horizontal coordinates	0.328	0.493
$Z_{up}$	upper crest position in vertical coordinates	0.083	0.124
$Z_{XX_{up}}$	upper crest curvature	-0.870	-0.580
$X_{lo}$	lower crest position in horizontal coordinates	0.325	0.488
$Z_{lo}$	lower crest position in vertical coordinates	-0.069	-0.046
$Z_{XX_{lo}}$	lower crest curvature	0.308	0.462
$\alpha_{te}$	trailing edge direction	-0.228	-0.152
$\beta_{te}$	trailing edge wedge angle	0.112	0.168

with re-interpolation should be used to seek the optimal solution.

#### A.2 Case A2 and A3: shape optimization of subsonic airfoil

The first and second aerodynamic cases (i.e., case A2 and A3) are the shape optimization of a subsonic airfoil with PARSEC method (Sobieczky 1999). PARSEC is an intuitive airfoil parameterization technique that has been widely used in a number of aerodynamic optimization studies. From the original 11 variables of PARSEC, we keep two parameters fixed (i.e., thickness and ordinate of trailing edge are set to zero) which leave us with nine variables as design variables. Figure 16 and Table 2 shows an illustration of PARSEC and the variables definition together with upper and lower bounds of the optimization problem used in this paper. We used the low-fidelity XFOIL (Drela 1989) code that couples a panel method and boundary layer solver to evaluate the aerodynamic performance for the subsonic case. Multiple independent runs can be evaluated due to the very fast evaluation times of XFOIL (i.e., less than one second). Based on a total of 160 independent runs from eight different methods, the highest  $L/D$  for case A2 and A3 is 164.5763 and 234.7368, respectively.

**Fig. 16** Illustration of PARSEC airfoil parameterization



**Table 3** The lower and upper bounds of the decision variables for transonic airfoil optimization problems (case A4 and A5)

$x$ (lower)	1	2	3	4	5	6	7	8
$l_b$	-0.153	-0.178	-0.131	-0.308	-0.11	-0.130	-0.06	0.05
$u_b$	-0.102	-0.118	-0.087	-0.205	-0.074	-0.087	-0.044	0.074
$x$ (lower)	9	10	11	12	13	14	15	16
$l_b$	0.102	0.103	0.141	0.097	0.190	0.133	0.159	0.164
$u_b$	0.153	0.154	0.212	0.146	0.285	0.199	0.239	0.247

### A.3 Case A4 and A5: shape optimization of an inviscid transonic airfoil

Case A4 and A5 deal with transonic airfoil optimization in inviscid flow. For this case, the airfoil is parameterized via class shape transformation (CST) (Kulfan 2008) that creates an aerodynamic shape by summing a number of Bernstein polynomials. The design variables according to the CST parameterization are the coefficients of these Bernstein polynomials. As a consequence, the CST parameters are not as intuitive as PARSEC; regardless, it is highly flexible since the number of parameters can be tuned. For cases A4 and A5, eight variables are used to create each the upper and lower surfaces (i.e., 16-variables in total). The upper and lower bounds for the CST design variables used in this paper are listed in Table 3. Based on a total of 160 independent runs from eight different methods, the highest  $L/D$  for case A4 and A5 are 303.86 and 290.59, respectively (note that this value is unrealistic for real-world design due to the use of Euler solver, nevertheless, our aim is to use these test cases so that we can perform statistical analyses).

### Appendix B: Test functions

1. F1: Branin function (two variables).

$$f_1(\mathbf{x}) = \left( b_2 - \frac{5.1}{4\pi^2} b_1^2 + \frac{5}{\pi} b_1 - 6 \right)^2 + 10 \left[ \left( 1 - \frac{1}{8\pi} \right) \cos(b_1) + 1 \right], \quad (23)$$

where  $b_1 = 15x_1 - 5$ ,  $b_2 = 15x_2$ , and  $x_1, x_2 \in [0, 1]^2$ .

2. F2: Sasena function (two variables).

$$f(\mathbf{x}) = 2 + 0.01(x_2 - x_1^2)^2 + (1 - x_1)^2 + 2(2 - x_2)^2 + 7\sin(0.5x_1)7\sin(0.7x_1x_2), \quad x_1 \in [0, 5], x_2 \in [0, 5]. \quad (24)$$

3. F3: Hosaki function (two variables)

$$y(\mathbf{x}) = \left( 1 - 8x_1 + 7x_1^2 - (7/3)x_1^3 + (1/4)x_1^4 \right) x_2^2 e^{-x_1}, \quad x_1 \in [0, 5], x_2 \in [0, 5]. \quad (25)$$

4. F4: Hartmann-3 function

$$f(\mathbf{x}) = - \sum_{i=1}^4 \alpha_i \exp \left( - \sum_{j=1}^3 A_{ij} (x_j - P_{ij})^2 \right). \quad \mathbf{x} \in (0, 1)^3. \quad (26)$$

5. F5: Hartmann-6 function

$$f(\mathbf{x}) = - \sum_{i=1}^4 \alpha_i \exp \left( - \sum_{j=1}^6 A_{ij} (x_j - P_{ij})^2 \right). \quad \mathbf{x} \in (0, 1)^6. \quad (27)$$

6. F6: Modified Rosenbrock function

$$f(\mathbf{x}) = \frac{1}{206} \left[ \sum_{i=1}^{k-1} 100(x_{i+1} - x_i^2)^2 + (1 - x_i^2 + \sum_{i=1}^k 75 \sin(5(1 - x_i)) - 300) \right], \quad \mathbf{x} \in [-1, 1]^5. \quad (28)$$

7. F7: Sphere function

$$f(\mathbf{x}) = \sum_i^k x_i^2, \quad \mathbf{x} \in [-5.12, 5.12]^9. \quad (29)$$

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### References

Acar E (2010) Various approaches for constructing an ensemble of metamodels using local measures. *Struct Multidiscip Optim* 42(6):879–896

Acar E (2013) Effects of the correlation model, the trend model, and the number of training points on the accuracy of Kriging metamodels. *Expert Syst* 30(5):418–428

Acar E, Rais-Rohani M (2009) Ensemble of metamodels with optimized weight factors. *Struct Multidiscip Optim* 37(3):279–294

Archambeau C, Bach F (2011) Multiple gaussian process models. [arXiv:1110.5238](https://arxiv.org/abs/1110.5238)

Bach FR (2009) Exploring large feature spaces with hierarchical multiple kernel learning. In: *Advances in neural information processing systems*, pp 105–112

Bader J, Zitzler E (2011) Hype An algorithm for fast hypervolume-based many-objective optimization. *Evol Comput* 19(1):45–76

- Bartoli N, Bouhlel M-A, Kurek I, Lafage R, Lefebvre T, Morlier J, Priem R, Stilz V, Regis R (2016) Improvement of efficient global optimization with application to aircraft wing design. In: 17th AIAA/ISSMO Multidisciplinary analysis and optimization conference, p 4001
- Ben Salem M, Tomaso L (2018) Automatic selection for general surrogate models. *Struct Multidisc Optim* 58(2):719–734
- Bishop CM (1995) *Neural networks for pattern recognition*. Oxford University Press
- Burnham KP, Anderson DR (2003) *Model selection and multimodel inference: a practical information-theoretic approach*. Springer
- Couckuyt I, Dhaene T, Demeester P (2014) oodace toolbox: a flexible object-oriented Kriging implementation. *J Mach Learn Res* 15:3183–3186
- Drela M (1989) Xfoil An analysis and design system for low Reynolds number airfoils. In: *Low Reynolds number aerodynamics*. Springer, pp 1–12
- Dubrulle O (1983) Cross validation of Kriging in a unique neighborhood. *J Int Assoc Math Geol* 15(6):687–699
- Durrande N, Ginsbourger D, Roustant O, Carraro L (2011) Additive covariance kernels for high-dimensional Gaussian process modeling. [arXiv:1111.6233](https://arxiv.org/abs/1111.6233)
- Forrester AI, Keane AJ, Bressloff NW (2006) Design and analysis of “noisy” computer experiments. *AIAA J* 44(10):2331–2339
- Ginsbourger D, Helbert C, Carraro L (2008) Discrete mixtures of kernels for Kriging-based optimization. *Qual Reliab Eng Int* 24(6):681–691
- Goel T, Haftka RT, Shyy W, Queipo NV (2007) Ensemble of surrogates. *Struct Multidiscip Optim* 33(3):199–216
- Hennig P, Schuler CJ (2012) Entropy search for information-efficient global optimization. *J Mach Learn Res* 13(Jun):1809–1837
- Jeong S, Murayama M, Yamamoto K (2005) Efficient optimization design method using Kriging model. *J Aircraft* 42(2):413–420
- Jones DR (2001) A taxonomy of global optimization methods based on response surfaces. *J Global Optim* 21(4):345–383
- Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. *J Global Optim* 13(4):455–492
- Kanazaki M, Matsuno T, Maeda K, Kawazoe H (2015) Efficient global optimization applied to wind tunnel evaluation-based optimization for improvement of flow control by plasma actuators. *Eng Optim* 47(9):1226–1242
- Kleijnen JP, van Beers W, Van Nieuwenhuysse I (2012) Expected improvement in efficient global optimization through bootstrapped Kriging. *J Global Optim*, 1–15
- Krige D (1951) *A statistical approach to some mine valuation and allied problems on the Witwatersrand*. By DG Krige
- Kulfan BM (2008) Universal parametric geometry representation method. *J Aircr* 45(1):142–158
- Liu H, Xu S, Wang X, Meng J, Yang S (2016) Optimal weighted pointwise ensemble of radial basis functions with different basis functions. *AIAA J*
- Marelli S, Sudret B (2014) Uqlab A framework for uncertainty quantification in matlab. In: *Vulnerability, uncertainty, and risk: quantification, mitigation, and management*, pp 2554–2563
- Martin JD, Simpson TW (2005) Use of Kriging models to approximate deterministic computer models. *AIAA J* 43(4):853–863
- Matheron G (1969) *Les cahiers du centre de morphologie mathématique de fontainebleau fascicule, vol 1. Le krigeage universel*. Ecole de Mines de, Paris, Fontainebleau
- Močkus J (1975) On Bayesian methods for seeking the extremum. In: *Optimization Techniques IFIP technical conference*. Springer, pp 400–404
- Mukhopadhyay T, Chakraborty S, Dey S, Adhikari S, Chowdhury R (2017) A critical assessment of Kriging model variants for high-fidelity uncertainty quantification in dynamics of composite shells. *Arch Comput Methods Eng* 24(3):495–518
- Namura N, Obayashi S, Jeong S (2016) Efficient global optimization of vortex generators on a supercritical infinite wing. *Journal of Aircraft*
- Palacios F, Alonso J, Duraisamy K, Colonno M, Hicken J, Aranake A, Campos A, Copeland S, Economon T, Lonkar A et al (2013) Stanford university unstructured (su 2): an open-source integrated computational environment for multi-physics simulation and design. In: *51st AIAA Aerospace Sciences meeting including the new horizons forum and aerospace exposition*, pp 287
- Picheny V, Wagner T, Ginsbourger D (2013) A benchmark of Kriging-based infill criteria for noisy optimization. *Struct Multidiscip Optim* 48(3):607–626
- Rasmussen CE, Williams CK (2006) *Gaussian processes for machine learning, vol 1*. MIT Press, Cambridge
- Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. *Statist Sci*, 409–423
- Samad A, Kim K-Y, Goel T, Haftka RT, Shyy W (2008) Multiple surrogate modeling for axial compressor blade shape optimization. *J Propuls Power* 24(2):302–310
- Shi H, Gao Y, Wang X (2010) Optimization of injection molding process parameters using integrated artificial neural network model and expected improvement function method. *Int J Adv Manuf Technol* 48(9–12):955–962
- Sóbester A, Leary SJ, Keane AJ (2004) A parallel updating scheme for approximating and optimizing high fidelity computer simulations. *Struct Multidiscip Optim* 27(5):371–383
- Sobieczky H (1999) *Parametric airfoils and wings*. In: *Recent development of aerodynamic design methodologies*. Springer, pp 71–87
- Stein ML (2012) *Interpolation of spatial data: some theory for Kriging*. Springer Science & Business Media
- Tesch M, Schneider J, Choset H (2011) Using response surfaces and expected improvement to optimize snake robot gait parameters. In: *2011 IEEE/RSJ International conference on intelligent robots and systems (IROS)*. IEEE, pp 1069–1074
- Viana FA, Haftka RT, Steffen V (2009) Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. *Struct Multidiscip Optim* 39(4):439–457
- Zhang J, Chowdhury S, Messac A (2012) An adaptive hybrid surrogate model. *Struct Multidiscip Optim* 46(2):223–238
- Zhou X, Ma Y, Tu Y, Feng Y (2013) Ensemble of surrogates for dual response surface modeling in robust parameter design. *Qual Reliab Eng Int* 29(2):173–197