Noname manuscript No.
(will be inserted by the editor)

Some Considerations Regarding the Use of Multi-fidelity Kriging in the Construction of Surrogate Models

David J.J. Toal

Received: date / Accepted: date

Abstract Surrogate models or metamodels are commonly used to exploit expensive computational simulations within a design optimization framework. The application of multi-fidelity surrogate modeling approaches has recently been gaining ground due to the potential for further reductions in simulation effort over single fidelity approaches. However, given a black box problem when exactly should a designer select a multi-fidelity approach over a single fidelity approach and vice versa? Using a series of analytical test functions and engineering design examples from the literature, the following paper illustrates the potential pitfalls of choosing one technique over the other without a careful consideration of the optimization problem at hand. These examples are then used to define and validate a set of guidelines for the creation of a multi-fidelity Kriging model. The resulting guidelines state that the different fidelity functions should be well correlated, that the amount of low fidelity data in the model should be greater than the amount of high fidelity data and that more than 10% and less than 80% of the total simulation budget should be spent on low fidelity simulations in order for the resulting multi-fidelity model to perform better than the equivalent costing high fidelity model.

Keywords Kriging · Multi-fidelity · Surrogate Modeling

D.J.J. Toal
Faculty of Engineering & the Environment
University of Southampton
University Road, Southampton, SO17 1BJ, U.K.
E-mail: djjt@soton.ac.uk

1 Introduction

Design optimization processes within a variety of industries often require the use of expensive computational simulations at their heart to determine a measure of the effectiveness or quality of a design. Such simulations can, in some instances, take several days to perform thereby ruling out the use of direct global optimization algorithms, such as genetic algorithms[8] or simulated annealing[15], within the optimization process. The use of surrogate modeling techniques within a design optimization loop, however, can dramatically reduce the number of actual simulations required and make the optimization process feasible.

Although there are a number of different surrogate modeling techniques[4, 19, 22], Kriging[16] is perhaps one of the most popular due to it's flexibility and the provision of a useful error metric. Since its initial application to the optimization of deterministic computational experiments by Sacks et al.[21], Kriging has grown in popularity and has been applied successfully to design problems in a variety of fields.

Kennedy and O'Hagan[14] extended the basic Kriging formulation to combine information from multiple levels of simulation fidelity into a more accurate surrogate model than would be created from employing only high fidelity data. As the performance of any surrogate based optimization is determined by the accuracy of the model a more accurate model can significantly reduce the total number of simulations required for an optimization. Such multi-fidelity approaches have been successfully employed throughout the literature in the design optimization of airfoils[17, 24, 26], wings[3], compressor rotors[2], combustors[25] and the creation of aerodynamic models[7, 9–11, 26].

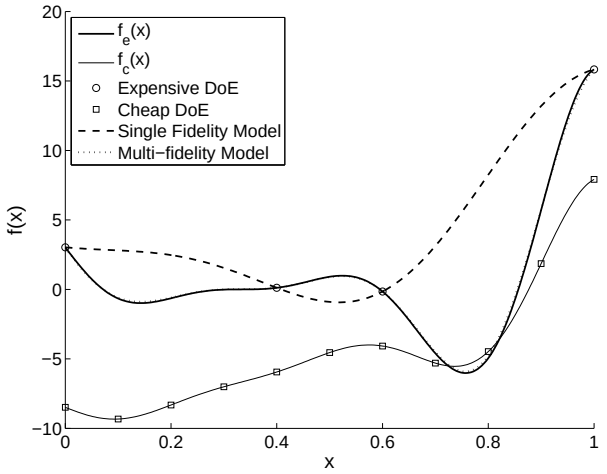


Fig. 1 An example of single and multi-fidelity Kriging[5]

Figure 1 is a simple example, recreated from Forrester et al.[5], of the advantages that multi-fidelity Kriging can offer if used to create a surrogate model. A Kriging model constructed from a four point design of experiments (DoE) of the expensive function ($f_e(x)$) is illustrated by the dashed line. Plainly this surrogate model does not represent the true function very well and any attempt to optimize using this model is hampered by this inaccurate prediction. Augmenting the four data points from the expensive function with an additional 11 data points from the 'cheap' function ($f_c(x)$) within a multi-fidelity Kriging model however, results in a very accurate model. In this case the surrogate model, represented by the dotted line, almost exactly matches the expensive function. Employing this surrogate within an optimization would greatly improve performance over the traditional Kriging model with the first update simulation being placed at almost the true global optimum.

Whilst this simple example illustrates the clear advantages that a multi-fidelity approach can bring, which has been mirrored by results presented within the literature[2,3,7,17,24], to date a number of important questions relating to the general application of such approaches have yet to be addressed:

1. Does the correlation between the low and high fidelity functions play a role in the effectiveness of a multi-fidelity Kriging model?
2. What role does the relative expense of the low and high fidelity functions play?
3. Does the total available budget of evaluations impact performance?
4. Given functions of similar cost what impact does the split between cheap and expensive evaluations have?
5. By considering these issues together is it possible to define a set of general guidelines for the use of a multi-fidelity Kriging model?

The following paper aims to investigate each of the above issues in turn and commences by briefly reviewing the formulation of both single and multi-fidelity Kriging models. The four analytical test functions used to investigate the above issues are then introduced. The impact of correlation between low and high fidelity functions and the magnitude of the cost ratio are then investigated. This is followed by an investigation into the effect of the total evaluation budget and the impact of the split between the number of low and high fidelity function evaluations for a fixed total budget. These investigations are then combined with additional results to produce a set of guidelines for effectively using multi-fidelity Kriging models. Finally, these guidelines are assessed with respect to three real life case studies taken from the literature, an engine SFC optimization, a compressor rotor optimization and a multi-point airfoil optimization.

2 Single & Multi-fidelity Kriging

The construction of a Kriging model is based upon the assumption that when two design points are close together the difference between their respective objective function values is small. This is modeled statistically by assuming that the correlation between two points,

$$\boldsymbol{R}_{ij} = \mathrm{Corr}\left[Y(\boldsymbol{x}_i), Y(\boldsymbol{x}_j)\right], \tag{1}$$

is given by,

$$\boldsymbol{R}_{ij} = \exp\left(-\sum_{l=1}^{d} 10^{\boldsymbol{\theta}^{(l)}} \|\boldsymbol{x}_i^{(l)} - \boldsymbol{x}_j^{(l)}\|^{\boldsymbol{p}^{(l)}}\right), \tag{2}$$

where $\boldsymbol{\theta}^{(l)}$ and $\boldsymbol{p}^{(l)}$ represent the, so called, hyperparameters of the l^{th} design variable. These hyperparameters are selected in order to maximize the likelihood on the observed dataset, $\boldsymbol{y}$, which equates to,

$$\phi = -\frac{n}{2}\ln(\hat{\sigma}^2) - \frac{1}{2}\ln(|\boldsymbol{R}|), \tag{3}$$

after simplification[12]. The equations,

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \mathbf{1}\hat{\mu})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu}), \tag{4}$$

and

$$\hat{\mu} = \frac{\mathbf{1}^T \boldsymbol{R}^{-1} \boldsymbol{y}}{\mathbf{1}^T \boldsymbol{R}^{-1} \mathbf{1}}, \tag{5}$$

give maximum likelihood estimates of the variance, $\hat{\sigma}^2$ and mean, $\hat{\mu}$, respectively, which can be used to calculate the likelihood function. In both the single and multi-fidelity Kriging models used here the hyperparameters are optimized using a hybridized particle swarm algorithm similar to that of Toal et al.[23].

With an optimal set of hyperparameters obtained the mean and the vector of correlations, $\boldsymbol{r}(\boldsymbol{x}^*)$, between an unknown point, $\boldsymbol{x}^*$ and the known sample points can be used to calculate the Kriging prediction,

$$y(\boldsymbol{x}^*) = \hat{\mu} + \boldsymbol{r}^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{1}\hat{\mu}). \tag{6}$$

Using the approach of Kennedy and O'Hagan[14] the high fidelity response is approximated by multiplying the low fidelity response by a scaling factor, ρ, and a Gaussian process representing the difference between the high and low fidelity data,

$$Z_e(\boldsymbol{x}) = \rho Z_c(\boldsymbol{x}) + Z_d(\boldsymbol{x}). \tag{7}$$

If $\boldsymbol{X}_e$ and $\boldsymbol{X}_c$ represent the expensive and cheap data respectively, then the covariance matrix $\boldsymbol{C}$ is,

$$\boldsymbol{C} = \begin{pmatrix} \sigma_c^2 \boldsymbol{R}_c(\boldsymbol{X}_c, \boldsymbol{X}_c) & \rho\sigma_c^2 \boldsymbol{R}_c(\boldsymbol{X}_c, \boldsymbol{X}_e) \\ \rho\sigma_c^2 \boldsymbol{R}_c(\boldsymbol{X}_e, \boldsymbol{X}_c) & \rho^2\sigma_c^2 \boldsymbol{R}_c(\boldsymbol{X}_e, \boldsymbol{X}_e) + \sigma_d^2 \boldsymbol{R}_d(\boldsymbol{X}_e, \boldsymbol{X}_e) \end{pmatrix} \tag{8}$$

where the correlations are of the same form as Eq. 2. Now, however, there are twice as many hyperparameters to determine, a set each for the Gaussian processes representing the cheap data and the difference between the cheap and expensive data and the scaling parameter, ρ.

As the low fidelity data and the differences between the low and high fidelity data are considered to be independent the hyperparameters defining the low fidelity Gaussian process can be determined in an identical manner to that of traditional Kriging. The hyperparameters defining the difference model are then determined by optimizing the log-likelihood as before, but using the difference between the cheap and expensive data,

$$\boldsymbol{d} = \boldsymbol{y}_e - \rho \boldsymbol{y}_c(\boldsymbol{X}_e), \tag{9}$$

instead of $\boldsymbol{y}$ in equations 3, 4 and 5. With the hyperparameters optimized the covariance matrix, Eq. 8, can be calculated and used in conjunction with a column vector, $\boldsymbol{c}$, of covariances of an unknown point to the known points to predict the high fidelity response at that unknown point,

$$y_e(\boldsymbol{x}^*) = \hat{\mu} + \boldsymbol{c}^T \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{1}\hat{\mu}), \tag{10}$$

where the mean is now given by,

$$\hat{\mu} = \frac{\boldsymbol{1}^T \boldsymbol{C}^{-1} \boldsymbol{Y}}{\boldsymbol{1}^T \boldsymbol{C}^{-1} \boldsymbol{1}}, \tag{11}$$

where $\boldsymbol{Y}$ is a combination of the known low and high fidelity responses, $\boldsymbol{Y}^T = [\boldsymbol{y}_c^T, \boldsymbol{y}_e^T]$. It should be noted that the upper and lower bounds on the θ and p hyperparameters are identical in both the single and multi-fidelity Kriging models with θ permitted to vary between -10 and 3 and p permitted to vary between 1.5 and 1.99. Note that the bounds of θ are equivalent to 1×10^{-10} and 100 respectively in the classical notation of Jones et al.[13] with the 10^{θ} term used in Eq. 2 to prevent values of 0 and improve the stability of the optimization. The scaling parameter ρ is permitted to vary between ± 5.

3 Analytical Test Functions

The Branin function is an analytical test function commonly used throughout the literature to test the performance of different surrogate modeling strategies. Here this function is the first of four such analytical functions used to test the performance of multi-fidelity Kriging under a variety of circumstances.

The traditional formulation of the Branin function,

$$f_e = (x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1 - \frac{1}{8\pi})\cos(x_1) + 10, \tag{12}$$

here plays the role of the response of an expensive, high fidelity simulation. Rather than having a single low fidelity response, as is the case in Figure 1, we consider a range of different low fidelity responses given by,

$$f_c = f_e - (A_1 + 0.5)(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2, \tag{13}$$

where the variable A_1 varies between 0 and 1 and effectively controls the level of correlation and error between the low and high fidelity responses. Figure 2 illustrates graphically the variation in both the r^2 correlation and the root mean square error (RMSE) as A_1 is varied where r^2 and RMSE are defined as

$$r^2 = \left(\frac{\sum_{i=1}^{n} (\boldsymbol{y}_{e_i} - \bar{\boldsymbol{y}}_e)(\boldsymbol{y}_{c_i} - \bar{\boldsymbol{y}}_c)}{\sqrt{\sum_{i=1}^{n} (\boldsymbol{y}_{e_i} - \bar{\boldsymbol{y}}_e)^2} \sqrt{\sum_{i=1}^{n} (\boldsymbol{y}_{c_i} - \bar{\boldsymbol{y}}_c)^2}} \right)^2 \tag{14}$$

and

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_{e_i} - \boldsymbol{y}_{c_i})^2} \tag{15}$$

respectively where $\boldsymbol{y}_e$ and $\boldsymbol{y}_c$ are a set of n observations of the expensive and cheap data for identical inputs with the bar denoting the mean of these sets. In this case the r^2 correlation varies from a maximum of 0.985 when $A_1 = 0$ to a minimum of approximately 3.8×10^{-4} when $A_1 = 0.514$.

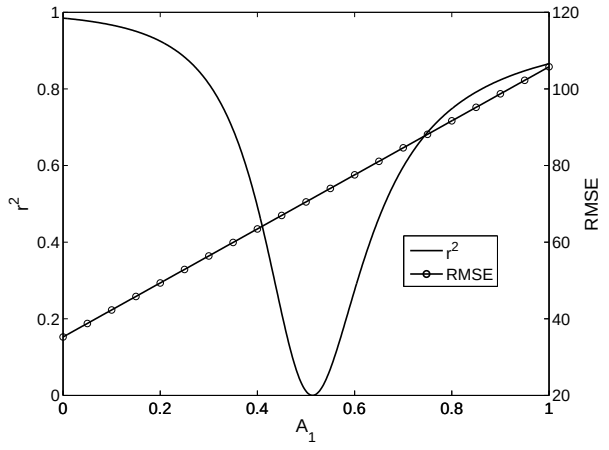


Fig. 2 Variation in r^2 and RMSE between the 'cheap' and 'expensive' Branin functions as A_1 varies

By varying A_1 and attempting to construct a multi-fidelity model using the resulting 'low' fidelity response and the Branin function, the impact of the correlation between the different fidelities on the accuracy of the resulting prediction can be ascertained. Interestingly, as A_1 is varied and the r^2 correlation increases beyond $A_1 = 0.514$ the RMSE continues to rise indicating that the cheap function is returning towards the general trend of the expensive function but with a considerable scaling error. Cases where $A_1 > 0.514$ therefore also enable the impact of RMSE between fidelity levels on multi-fidelity surrogate model accuracy to be examined.

The second analytical test function considered is the Paciorek function described by,

$$f_e = \sin\left(\frac{1}{x_1 x_2}\right), \tag{16}$$

where the 'cheap' version of this function is defined by,

$$f_c = f_e - 9A_2^2 \cos\left(\frac{1}{x_1 x_2}\right), \tag{17}$$

with the parameter A_2 permitted to vary between 0 and 1 and causing the variation in r^2 correlation and RMSE between f_e and f_c shown in Figure 3.

The third analytical test function considered here is the three variable Hartmann H34 function defined by,

$$f_e = -\sum_{i=1}^{4} \alpha_i \exp\left[-\sum_{j=1}^{3} \beta_{ij}(x_j - P_{ij})^2\right], \tag{18}$$

where,

$$\alpha = \begin{bmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{bmatrix} \beta = \begin{bmatrix} 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{bmatrix} P = \begin{bmatrix} 3689 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{bmatrix} \times 10^{-4},$$

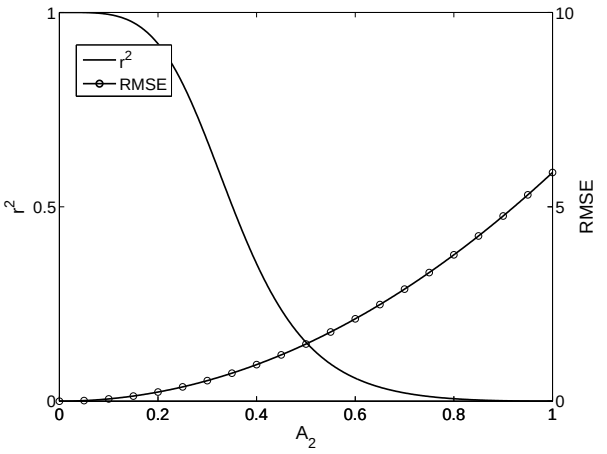


Fig. 3 Variation in r^2 and RMSE between the 'cheap' and 'expensive' Paciorek functions as A_2 varies

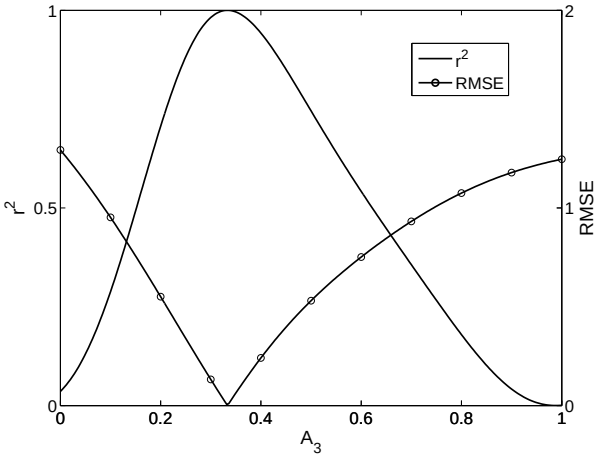


Fig. 4 Variation in r^2 and RMSE between the 'cheap' and 'expensive' Hartmann H34 functions as A_3 varies

and the associated parametric 'cheap' function is given by,

$$f_c = -\sum_{i=1}^{4} \alpha_i \exp\left[-\sum_{j=1}^{3} \beta_{ij}(x_j - \frac{3}{4}P_{ij}(A_3 + 1))^2\right], \tag{19}$$

where by varying A_3 from 0 to 1 the variation in r^2 and RMSE shown in Figure 4 is achieved.

The fourth and final analytical test function is the 10 variable Trid function defined by,

$$f_e = \sum_{i=1}^{10}(x_i - 1)^2 - \sum_{i=2}^{10} x_i x_{i-1}, \tag{20}$$

where, $x_i \in [-100, 100]$, and the associated 'cheap' parametric function is given by,

$$f_c = \sum_{i=1}^{10}(x_i - A_4)^2 - (A_4 - 0.65)\sum_{i=2}^{10} i x_i x_{i-1}. \tag{21}$$

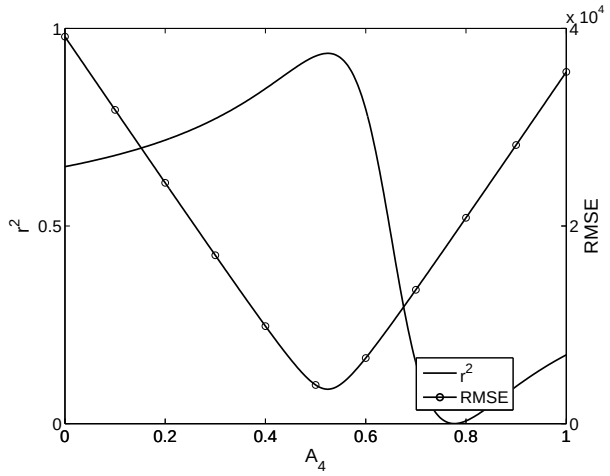


Fig. 5 Variation in r^2 and RMSE between the 'cheap' and 'expensive' Trid functions as A_4 varies

As with the previous analytical test functions the parameter A_4 varies between 0 and 1 resulting in the variation in r^2 and RMSE illustrated in Figure 5.

As was the case with the Branin function the parametric 'cheap' Trid function permits cases where there exist relatively high correlations and high errors between the fidelity levels to be investigated (when $A_4 < 0.4$ in Figure 5). This parametric function also further permits the impact of RMSE to be investigated by including instances where the error between the functions is relatively small but the correlation is low, i.e. when $0.7 < A_4 < 0.8$.

4 The Impact of Function Correlation & Cost Ratio

Given a set of parametric analytical test functions, let us now utilise these models to investigate the impact of both the level of correlation between the cheap and expensive functions and the cost ratio on the performance of a multi-fidelity Kriging model. In this investigation the four A parameters are varied for each function and a multi-fidelity surrogate model is constructed using a variety of function evaluation cost ratios and compared to a single fidelity Kriging model of equivalent cost.

In all cases the multi-fidelity models are compared to a single fidelity model constructed from $5d$ sample points, where d is the number of dimensions in the underlying problem. In the case of the Branin, Paciorek, Hartmann H34 and Trid functions this equates to a total of 10, 10, 15 and 50 sample points respectively.

The multi-fidelity surrogate models are constructed by replacing d 'expensive' function evaluations with 'cheap' evaluations. That is to say that $4d$ 'expensive' sample points are used in each multi-fidelity model. In the case of the Branin, Paciorek, Hartmann H34 and Trid functions this equates to a total of 8, 8, 12 and 40 'expensive' sample points respectively.

The total number of cheap sample points is then defined by multiplying the number of expensive points replaced, d, by the cost ratio of the expensive to cheap functions. A cost ratio of 4:1, for example, indicates that an evaluation of the cheap function is assumed to be one quarter the cost of an evaluation of the expensive function. The total number of cheap evaluations in a multi-fidelity surrogate employing such a ratio is therefore $4d$.

To help illustrate this more clearly lets consider a few simple examples. As noted above, the single fidelity model of the Branin function is assumed to have 10 expensive sample points. Assuming a cost ratio of 4:1 therefore means that the single fidelity model is compared to a multi-fidelity model consisting of 8 expensive sample points and $4d = 8$ cheap sample points. In the case of a 15:1 cost ratio an evaluation of the cheap function is assumed to be one fifteenth the cost of an expensive function evaluation. The multi-fidelity surrogate model in this instance will consist of 8 expensive sample points, as before, but these are now augmented by $15d = 30$ cheap function evaluations. Extending this to the ten dimensional Trid function, the single fidelity model will consist of 50 expensive function evaluations whereas a multi-fidelity model constructed, assuming a 10:1 cost ratio, will consist of 40 expensive function evaluations and $10d = 100$ cheap function evaluations.

For both the single and multi-fidelity cases a random Latin-Hypercube sampling plan is used to define the sample points from which the surrogate models are constructed. In the case of the multi-fidelity surrogate models an initial large sampling plan is constructed for the cheap function with an optimal space filling sub-set of this sampling plan defined using a max-min criteria[6]. This optimal sub-set is then evaluated using the expensive test function.

Both the single and multi-fidelity Kriging models, once constructed, are assessed for accuracy using a set of test points evaluated from the true high fidelity function. These test points are separate to the sampling plans used to construct the surrogate. In the case of the Branin and Paciorek functions 1000 test points are used while 5000 points are used for the Hartmann H34 function and 10,000 for the Trid function. With the surrogate model predictions at these points determined the r^2 correlation and RMSE of the prediction is calculated. To mitigate the impact of the sampling plan the results are averaged over 50 different sampling plans for the Branin, Paciorek and Hartmann H34 functions and

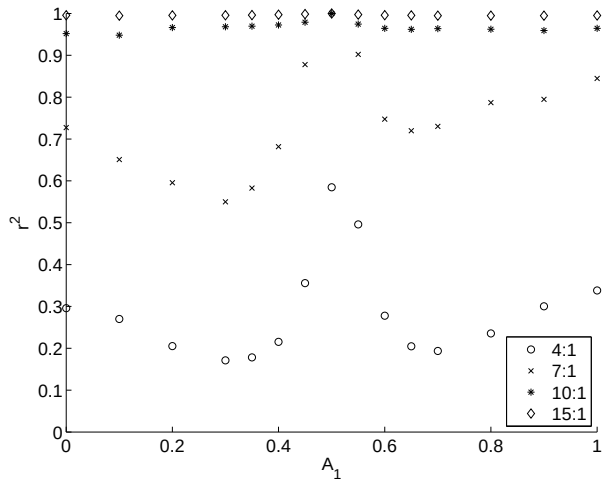


Fig. 7 r^2 correlation between the low fidelity Kriging prediction and the cheap Branin function

over 25 different plans for the Trid function due to the expense of constructing a 10 dimensional model.

Figures 6(a) and 6(b) illustrate the variation in both the r^2 correlation and the RMSE of the multi-fidelity prediction of the Branin function as A_1 is varied for a number of different cost ratios. The dashed line of Figures 6(a) and 6(b) illustrate the accuracy achieved by the baseline single fidelity Kriging approach using a 10 point Latin-Hypercube sampling plan. Figure 7 illustrates the prediction accuracy of the Kriging model describing the low fidelity function with varying A_1 and function cost ratio.

Comparing Figures 6(a) and 6(b) to Figure 2 the impact of the correlation between the low and high fidelity functions for the Branin function is immediately obvious. The results presented in Figures 6(a) and 6(b) show a clear trend whereby those cases with a much better correlation tend to produce more accurate multi-fidelity surrogate models. For those cases where the r^2 correlation between the different function fidelities is above approximately 0.6 the resulting multi-fidelity model can be more accurate than the single fidelity model. We can also observe from Figure 6 the tendency for the performance of the multi-fidelity model to closely match the correlation between the two fidelity levels and not the RMSE. For the case when $A_1 = 1.0$ the correlation between the two functions is high and so to is the RMSE, however, the multi-fidelity model is considerably more accurate than the single fidelity model for this case.

The results presented in Figures 6(a) and 6(b) also demonstrate that the number of cheap simulations used in the construction of a multi-fidelity model also has a clear impact on its accuracy. With a cost ratio of 4:1 there are as many cheap as expensive simulations within the surrogate and the performance is extremely poor with the r^2 in the majority of cases half that of the baseline Kriging model. Increasing the cost ratio to 7:1 increases the amount of low fidelity data to 14 points and greatly improves the accuracy of the resulting model so that in some instances it outperforms the baseline Kriging model.

Generally performance continues to improve as more low fidelity data is included. This improvement in performance begins to plateau when a total of 30 low fidelity data points are included. This indicates that multi-fidelity Kriging performance is dependent on the accuracy of the Gaussian process representing the low fidelity model. Within a multi-fidelity Kriging model the low fidelity model helps to guide the high fidelity model in regions where there is no high fidelity data available, the more accurate this model the better it can guide the high fidelity data and the presented results confirm this.

This is further confirmed when one considers the accuracy to which the low fidelity response is represented by the low fidelity Gaussian process, Figure 7. For the case when $A_1 = 0$ where f_c is highly correlated with f_e and for an assumed cost ratio of 4:1 the mean r^2 was 0.296 with a mean RMSE of 25.170. The 15:1 case, with 30 cheap data points, represents the true response of the cheap model much more accurately with a mean r^2 of 0.996 and a mean RMSE of 1.584 hence we observe a corresponding improvement in the accuracy of the multi-fidelity Kriging model in Figure 6.

Setting $A_1 = 0.5$ so that the correlation between f_c and f_e is very low, $r^2 \approx 0$, provides an interesting case. Here the 4:1 case is better able to represent the low fidelity response with a mean r^2 of 0.585 and a mean RMSE of 4.229 likewise the 15:1 case is also more accurate with a mean r^2 of 1.000 and a mean RMSE of 6.1×10^{-3}. However, even with such a large difference in surrogate accuracy there is very little improvement in the quality of the multi-fidelity prediction for these two cases. This indicates that for cases with a poor correlation between cheap and expensive data additional cheap information may not help improve overall accuracy.

Figures 8(a) and 8(b) illustrate the variation in the accuracy of a multi-fidelity Kriging model as the number of 'cheap' data points varies and with varying A_2 for the Paciorek function. Figure 9 illustrates the accuracy with which the low fidelity Kriging model is constructed with varying cost ratio for different values of A_2. As with the Branin function the results of Figure 8 also demonstrate that the overall accuracy of the multi-fidelity model is dependent on both the correlation between the cheap and expensive functions and the

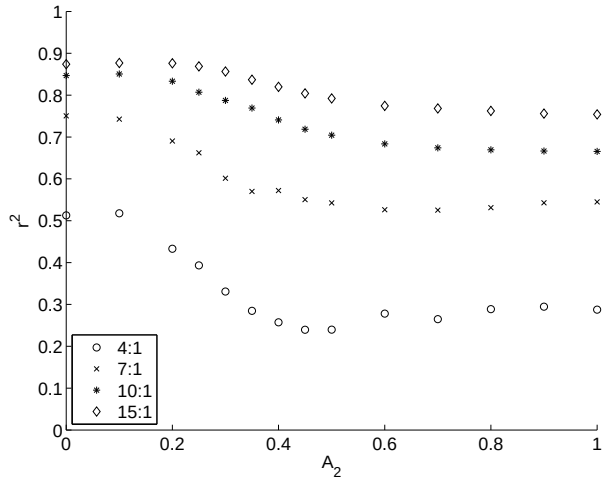


Fig. 9 r^2 correlation between the low fidelity Kriging prediction and the cheap Paciorek function

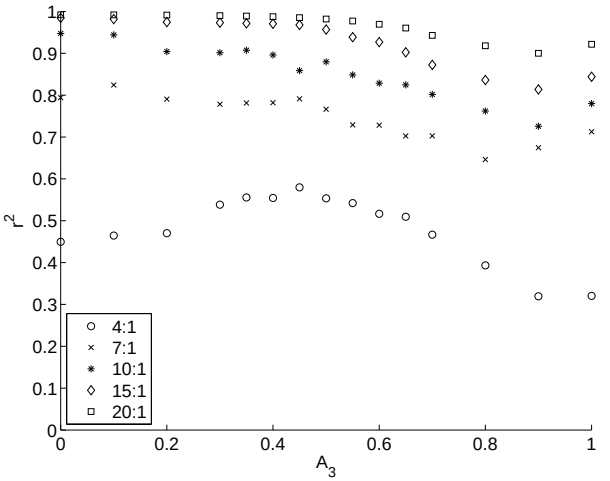


Fig. 11 r^2 correlation between the low fidelity Kriging prediction and the cheap Hartmann H34 function

amount of low fidelity data which controls the accuracy to which the underlying cheap function is represented.

However, unlike the Branin function where the additional accuracy of the cheap response improved the quality of models created using poorly correlated functions, the results here indicate that the improvement in the accuracy of the cheap function can actually exacerbate a reduction in the accuracy of the multi-fidelity model. For the case when $A_2 = 1$ and the true r^2 is approximately 0 an increase in the number of data points improves the accuracy of the cheap model from a r^2 of 0.289 and RMSE of 4.512 when there are 8 cheap points to a r^2 of 0.754 and RMSE of 2.597 when there are 30 points, see Figure 8. But as illustrated in Figures 8(a) and 8(b) there is a reduction in r^2 and an increase in RMSE of the resulting multi-fidelity model when the number of data points is increased.

Figures 10(a) and 10(b) illustrate the variation in the accuracy of a multi-fidelity model of the Hartmann H34 function when the correlation between functions as well as the cost ratio and therefore the number of cheap data points varies. Figure 11 illustrates the accuracy with which the low fidelity surrogate model represents the low fidelity function with varying A_3 and cost ratio. Unlike the Branin and Paciorek functions, in all cases there are now a total of 12 expensive data points used with the equivalent of three expensive points converted into cheap data evaluations. The 15:1 cost ratio case therefore employs 45 cheap and 12 expensive data points. Once again the accuracy of each multi-fidelity model is compared to the accuracy of the equivalent Kriging model, which, in this case employs 15 data points.

As for the previous cases, both the correlation between the cheap and expensive functions and the number of data points used to represent the cheap response play an important role in the accuracy of the multi-fidelity model. As was also demonstrated by the Paciorek test function, the performance of those cases where there is a poor correlation actually degrades as the accuracy of the cheap function is increased.

For the case when $A_3 = 0.35$ and the cost ratio is 4:1 and there are 12 cheap data points, the cheap model has a r^2 of 0.556 and a RMSE of 0.685. Increasing the number of points to 60, as in the 20:1 case, improves the accuracy of the cheap response with the r^2 now 0.989 and the RMSE now 0.101. As clearly illustrated in Figures 10(a) and 10(b) there is a corresponding improvement in the accuracy of the final multi-fidelity model. However, for the case when $A_3 = 1$ and the functions are poorly correlated, whilst the accuracy of the cheap model also improves with the r^2 increasing from 0.320 to 0.922 and the RMSE decreasing from 0.211 to 0.060 the accuracy of the resulting multi-fidelity model noticeably reduces.

Figures 12(a) and 12(b) illustrate the variation in the accuracy of a multi-fidelity model when the number of cheap data points and correlation between the cheap and expensive function varies varies for the Trid function. While Figure 13 illustrates the accuracy with which the low fidelity function is recreated by the low fidelity surrogate model with varying cost ratio and A_4. In this case the benchmark single fidelity Kriging model is constructed using a total of 50 expensive sample points while each of the multi-fidelity models are constructed from 40 expensive sample points with the remaining budget of 10 expensive points converted into cheap sample points according to the defined cost ratio. A surrogate model constructed using the 4:1 cost ratio therefore consists of 40 expensive sample points

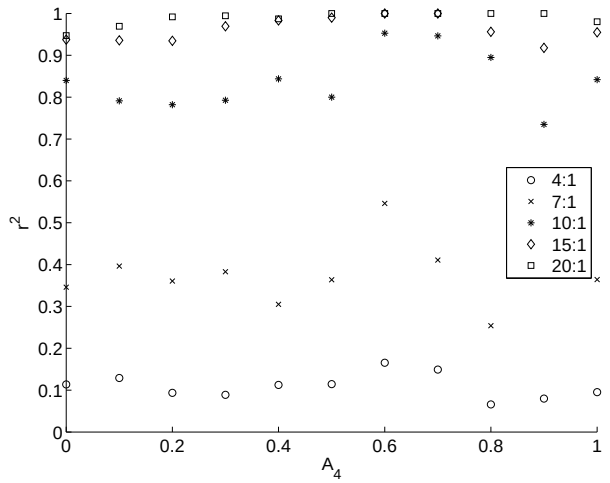


Fig. 13 r^2 correlation between the low fidelity Kriging prediction and the cheap Trid function

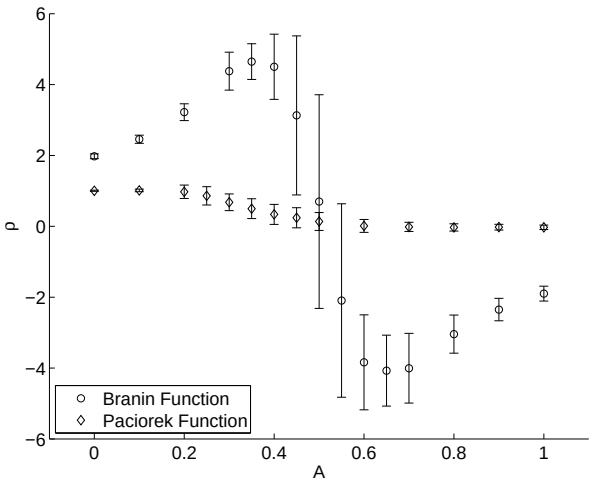


Fig. 14 Variation in ρ with A_1 and A_2 for the Branin and Paciorek functions

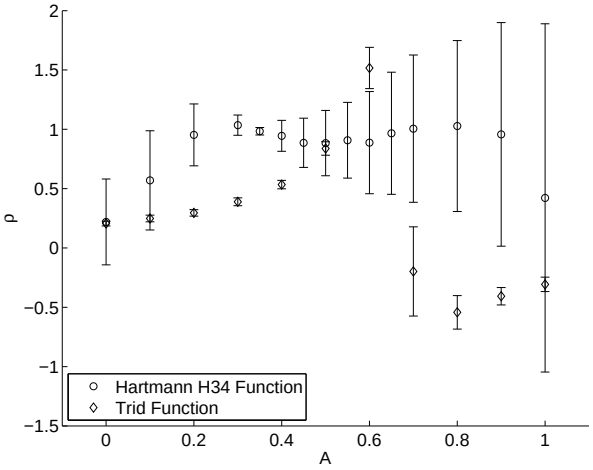


Fig. 15 Variation in ρ with A_3 and A_4 for the Hartmann H34 and Trid functions

and 40 cheap sample points while a model constructed using the 20:1 ratio employs 40 expensive and 200 cheap sample points.

The trends illustrated by Figures 12(a) and 12(b) are very similar to those observed for the previous three test functions. The performance of the multi-fidelity model closely follows the level of correlation between the cheap and expensive functions with the number of cheap simulations and hence the accuracy of the cheap model also impacting the performance.

As stated in equation 9 the multi-fidelity model is constructed from a low fidelity model multiplied by a scaling factor and added to a second model of the difference between the low and high fidelity data. The results presented above seem to suggest that even when the low fidelity model is quite accurate the log-likelihood optimization of the hyperparameters, which includes the scaling factor ρ, is producing surrogate models where the low fidelity model appears to be more important than it should be. To investigate this further, consider therefore the variation in ρ for the cases presented previously where a large amount of cheap data is available. For the Branin and Paciorek functions, this is when a 15:1 cost ratio is used while for the H34 and Trid functions this is when a 20:1 cost ratio is used. Considering only these cases reduces the impact of an inaccurate low fidelity surrogate model on the analysis of ρ.

Figures 14 and 15 illustrate the mean and standard deviation in the magnitude of ρ as the A parameters are varied for all four test functions. Comparing these figures to those of the correlations between the low and high fidelity versions of each analytical function a number of trends can be observed. As suggested above, in the majority of cases the optimum value of ρ determined through the likelihood optimization does not approach zero. For the Branin, Hartmann H34 and Trid function when there is little correlation between the low and high fidelity functions the magnitude of ρ can be significant. Only with the Paciorek function does the magnitude of ρ appear to tend towards zero as the correlation reduces. The second observable trend is the tendency for there to be a much higher variation in the magnitude of ρ in cases with very low correlation between the fidelity levels. In the case of the Branin and Hartmann H34 functions there is a considerable spread in the values of ρ resulting from the hyperparameter optimization when the functions are poorly correlated. These observations suggest that the above multi-fidelity Kriging formulation takes little notice of the correlation between functions and can tend put emphasis on the low fidelity model when it should not.

The above investigations point to a number of interesting conclusions. Firstly the level of correlation be-

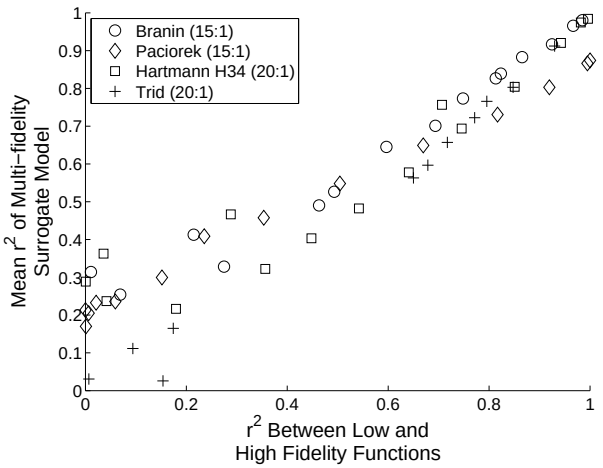


Fig. 16 Plot of mean r^2 correlation of the multi-fidelity prediction against the r^2 correlation between the low and high fidelity functions for cases with large amounts of low fidelity data

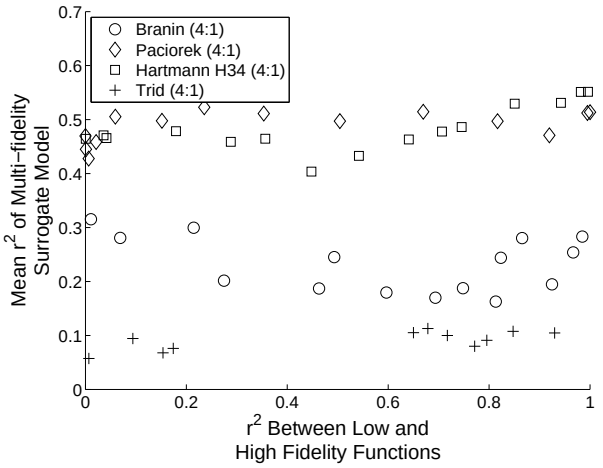


Fig. 17 Plot of mean r^2 correlation of the multi-fidelity prediction against the r^2 correlation between the low and high fidelity functions for cases with little low fidelity data

tween the low and high fidelity functions defining a multi-fidelity model is extremely important in determining the accuracy of the model. This is reinforced via Figure 16 which plots the mean r^2 correlation resulting from a multi-fidelity model against the r^2 correlation between the low and high fidelity function for all four test functions. This figure demonstrates a clear link between the function correlation and the accuracy of the model. Secondly, a high number of cheap data points is also beneficial to the accuracy of the multi-fidelity model but only for closely correlated functions. Figure 17 is a recreation of Figure 16 but for those instances when there is very little low fidelity data. This plot clearly illustrates much less of a correlation between the correlation of the two fidelities and the accuracy of the resulting multi-fidelity model. In some instances for badly correlated functions a large number of cheap data points can actually have a negative impact on model accuracy. The above results also indicate that there should always be more cheap data points than expensive, that the formulation of a multi-fidelity Kriging model can put emphasis on inappropriate low fidelity models and that the RMSE between the fidelity levels plays much less of a role compared to the correlation between the fidelities.

5 The Impact of the Number of Expensive Evaluations

An alternative way of increasing the accuracy of any surrogate model is to increase the amount of data used in its construction. Increasing the amount of high fidelity data within a multi-fidelity Kriging model therefore increases the amount of data used to construct the difference model. In theory the more data within this model the more accurate the difference model and the better the multi-fidelity predictor, as a whole, can overcome the differences in the correlation between the cheap and expensive data.

Figures 18(a) and 18(b) illustrate the variation in both r^2 and RMSE as the number of expensive function evaluations is increased for the Branin function. In all cases 30 cheap evaluations form the basis of the model with an optimal expensive subset selected from it. As before the accuracy of each model is averaged over 50 different sampling plans.

These figures illustrate that more expensive data does indeed improve the quality of the model even when constructed between considerably uncorrelated functions. For the case where $A_1 = 0.5$ and the correlation between the low and high fidelity functions is approximately zero, as more expensive data is added both the r^2 correlation and RMSE of the resulting model improve considerably.

Of course, as the amount of high fidelity data is increased so too is the amount of data which can be used in an equivalent costing single fidelity model. A series of Kriging models were therefore constructed using the number of expensive function evaluations used in the multi-fidelity Kriging model plus an additional two. A multi-fidelity model with 18 expensive evaluations is therefore compared to a single fidelity model constructed from 20.

Both Figures 18(a) and 18(b) include a dashed line representing the point at which the multi-fidelity models perform better than their equivalent costing Kriging models. Those models inside of this dashed line are less accurate than the Kriging model while those outside of

this line are more accurate. Clearly, even though the accuracy of the multi-fidelity model improves as more expensive data is employed the Kriging model is still performing better for a significant range of A_1 values. In this case for the range $0.33 < A_1 < 0.7$, or when the correlation between the functions is less than 0.8, no matter how many cheap or expensive function evaluations are employed in the multi-fidelity model the single fidelity model will be more accurate. In other words, for these cases the addition of cheap data actually misleads the predictor reducing its accuracy.

Figures 19(a) and 19(b) illustrate the variation in the accuracy of the multi-fidelity model when the number of expensive data points is varied but a constant 30 cheap data points are employed for the Paciorek function.

As with the results for the Branin function, the increase in the amount of expensive data improves the accuracy of the model, even when poorly correlated data is employed. However, when compared to an equivalent costing single fidelity Kriging model, as with the Branin function, there is a clearly defined region where the multi-fidelity model performs better than Kriging no matter the number of expensive data points used. For the Pacoriek function this occurs when $A_2 < 0.35$ or when the correlation between the low and high fidelity functions is greater than approximately 0.5.

Figure 20 illustrates the change in both r^2 correlation and RMSE as the amount of expensive data is increased for the Hartmann H34 function. In this case a cost ratio of 20:1 is assumed throughout with three expensive points assumed to be sacrificed for 60 cheap data points. A multi-fidelity model with 60 cheap and 15 expensive data points is therefore compared to a single fidelity model with 18 expensive data points.

As with both the Branin and Paciorek functions, increasing the amount of expensive data improves the accuracy of the multi-fidelity model no matter the correlation between the cheap and expensive data. However, when the accuracy of these models is compared to the equivalent costing Kriging model there is a clear region where the Kriging model performs better. In this case the multi-fidelity model is only more accurate between the two dotted lines equating to $0.17 < A_3 < 0.55$ which equates the region where the correlation between the two functions is greater than 0.53.

Figure 21 illustrates the change in both r^2 correlation and RMSE as the amount of expensive data is increased for the Trid function. In this case 10 expensive simulations have been replaced by 150 cheap simulations. The 15:1 cost ratio is used as Figure 12 illustrated very little difference in performance between a cost ratio of 15:1 and 20:1. In Figure 21 the performance of a Kriging model constructed from 40 expensive function evaluations is therefore compared to a multi-fidelity model with 30 expensive and 150 cheap function evaluations.

The dashed lines of Figure 21 indicate a clear region, when $A_4 < 0.6$, or when the correlation between the low and high fidelity functions is greater than 0.72, where the multi-fidelity model out performs the equivalent costing single fidelity model. As with the other test functions the more expensive data points that are included in the model the more accurate the model generally becomes. However, it is only when correlation between the cheap and expensive functions is greater than 0.72 that in general the multi-fidelity model is more accurate than the Kriging model.

The dashed line in the top left corner of Figure 21 indicates a region where the Kriging model outperforms the expensive model even though Figure 5 indicates that the functions are very well correlated. Even though the mean r^2 of a Kriging model with 100 sample points is 0.72, higher than the multi-fidelity model in this region, the consistency in the quality of the model is greatly reduced. The standard deviation of the r^2 of the equivalent multi-fidelity model when $A_4 = 0.0$ is 0.12 whereas that for the Kriging model is 0.35, a considerable increase.

From the above investigation into the impact of the number of expensive function evaluations it is clear that increasing the amount of expensive data improves the accuracy of any multi-fidelity surrogate model irrespective of the level of correlation between the low and high fidelity functions. However, an equivalent costing single fidelity surrogate model will still perform better than a multi-fidelity model if the correlation between the underlying functions is low. It is interesting to note that the bounds of the regions illustrated in Figures 18, 19, 20 and 21 where the multi-fidelity model performs better are generally along lines of constant A i.e. a constant correlation between the underlying functions. This suggests that when such a surrogate model is employed in a cyclic process where additional infill points are generated and included within the model, as is the case in an optimization, an underlying poor correlation will always put the multi-fidelity model at a disadvantage no matter how many additional points are added.

6 The Impact of the Ratio of Expensive to Cheap Evaluations

The results presented in Figures 18, 19, 20 and 21 assumed that the low fidelity function was considerably less expensive than the high fidelity function. Of course in reality the costs of these functions may be relatively

similar. Consider now the worst case scenario for each of the above analytical functions where the cheap function is only half the cost of the expensive function.

As already indicated above, the results of Figures 6, 8, 10 and 12 demonstrated that there must always be more cheap data than expensive. For the test functions considered above and assuming that the total evaluation budget remains $5d$ this somewhat constrains the number of potential ways in which the simulation budget can be split up when the cost ratio is 2:1. For the Branin and Paciorek functions at most, six high fidelity evaluations can be combined with eight low fidelity evaluations, for the Hartmann H34 this means that at most nine high fidelity evaluations can be combined with 12 low fidelity evaluations and for the Trid function 33 high fidelity evaluations can be combined with 34 low fidelity.

Figure 22 is a recreation of Figure 6 with a fixed 2:1 cost ratio and different splits between the number of cheap and expensive function evaluations for the Branin function. The legend of both graphs indicate the number of cheap and expensive evaluations respectively and as per the previous investigations all results are averaged over 50 different sampling plans.

As per the results of Figure 6, Figure 22 clearly indicates the importance of having a large amount of low fidelity information in the multi-fidelity model. The more cheap data the more accurate the low fidelity surrogate and the better it can guide the high fidelity prediction.

However, unlike Figure 6, Figure 22 better illustrates the pitfalls of creating a multi-fidelity model using low and high fidelity objective functions of similar costs. In such a case it is even more important for the low and high fidelity functions to be well correlated. Only then will it be worthwhile foregoing high fidelity function evaluations for low fidelity function evaluations.

Even if the functions are closely correlated Figure 22 also indicates that it's very important to get the split between cheap and expensive evaluations correct. Figure 22 suggests that a split of three expensive and 14 cheap function evaluations will outperform the equivalent costing single fidelity model if the r^2 correlation between the two functions is greater than approximately 0.95.

Figure 23 illustrates the variation in the accuracy of multi-fidelity models constructed for the Paciorek function using a fixed 2:1 cost ratio when different numbers of expensive function evaluations are used. As with the Branin function, the assumption of similar costing objective functions makes it extremely important for the cheap and expensive functions to be well correlated. Figure 8 indicates that only when the r^2 correlation is above 0.9 is there any advantage to employing a multi-fidelity model and even then the split between the number of expensive and cheap function evaluations must be carefully considered. As with the Branin function three expensive function evaluations in combination with 14 cheap function evaluations performs best out of those strategies considered here.

Figure 24 illustrates a similar trend for the Hartmann H34 function. As with the Branin and Paciorek functions, these illustrate the importance of highly correlated cheap and expensive functions when the costs of those functions are relatively similar.

Only when the r^2 correlation is greater than approximately 0.95 are both the RMSE and r^2 correlation of the resulting multi-fidelity prediction better than the equivalent costing single fidelity model. Similarly, the split between the number of expensive and cheap function evaluations plays an important role with the strategy with the smallest number of expensive evaluations performing best.

The Trid function, Figure 25, also illustrates the need for a careful consideration of the split between the expensive and cheap functions and the importance of the close correlation of these functions when they are of a similar cost. As with the previous example it is only those cases where the functions are highly correlated and there is a large amount of cheap data that perform better than the equivalent costing Kriging model.

The results for these four test functions therefore tend to suggest that when the functions are of a similar cost the level of correlation plays even more of a role in the accuracy of the resulting multi-fidelity model and that in these cases it's important to be well correlated thereby permitting to use of a much smaller number of expensive function evaluations.

7 Derivation of a Best Practice

The previous investigations have investigated the impact of function correlation, total evaluation budget, evaluation cost ratio and the split in the global evaluation budget between cheap and expensive simulations. However, it could be argued that these factors are themselves interlinked and to develop a more general set of best practice rules one should consider the simultaneous impact of each of these aspects on the creation of a multi-fidelity model.

Towards that end let us now consider the performance of a multi-fidelity Kriging model constructed for each of the above analytical test functions but simultaneously taking into account the level of function correlation, r^2 the relative expense of the cheap function

evaluation, C_r, and the fraction of the expensive functions replaced by cheap functions, f_r. By relative expense we mean,

$$C_r = \frac{C_c}{C_e}, \tag{22}$$

where C_c is the cost of a cheap function evaluation and C_e is the cost of an expensive function evaluation. A $C_r = 0.1$, for example, indicates that an expensive evaluation is 10 times the cost of a cheap evaluation. The fraction of expensive function evaluations replaced by cheap evaluations, f_r, is defined as,

$$f_r = 1 - \frac{n_{me}}{n_{se}}, \tag{23}$$

where n_{me} is the number of expensive evaluations in the multi-fidelity model and n_{se} is the number of expensive evaluations in the equivalent costing single fidelity model. If a total of 10 expensive simulations can be afforded then an f_r of 0.8 means that two of these will be replaced by cheap evaluations of equivalent cost where the number of cheap evaluations is then dependent on the cost ratio C_r.

The previous investigations illustrated that the impact of the level of function correlation with increasing numbers of expensive evaluations is relatively constant therefore allowing us to ignore it in the following study. This leaves, what is essentially, a three dimensional hypercube of potential multi-fidelity Kriging settings with the previous studies presented in Sections 4 and 6 forming lines through this space.

For each test function let us perform what is essentially a full factorial sampling plan within this hypercube of settings. For both the Branin and Paciorek functions 10-90% of the total evaluation budget will be replaced with cheap data. For each of these settings, cases will be run for which the cheap simulations are assumed to cost 1/2, 1/3, 1/4, 1/7, 1/10 and 1/15 times the cost of an expensive evaluation. For each of these cases the A parameters are varied in an identical manner to that of Figures 6 and 8 thereby adjusting the correlation between the low and high fidelity functions. As with the previous cases the results are averaged over 50 different sampling plans and a total budget of $5d$ sample points is assumed. The Branin and Paciorek functions will therefore have a full factor sampling plan within the settings hypercube of 720 and 672 points respectively. A similar process is carried out for both the Hartmann H34 and Trid function but as there are more sample points in both of these cases it is possible to consider a much wider range of percentages of the total evaluation budget replaced with cheap simulations. For the Hartmann H34 function 14 different percentages are considered ranging from 6.7% to 93%. In the case of the Trid function an even greater range of replacement percentages is investigated from 6% to 94%. The relative simulation costs considered for both of these functions are the same as those used for the Branin and Paciorek functions with the A parameters varying in an identical manner to the previous studies.

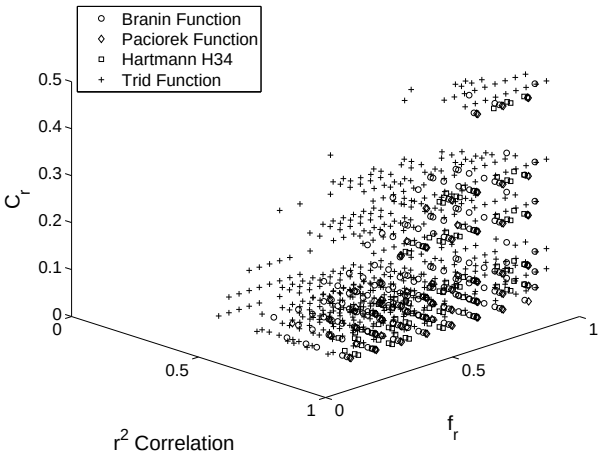


Fig. 26 Overview of settings for better multi-fidelity than single fidelity performance for all four test functions.

With the average performance of the multi-fidelity surrogate model at each of these conditions calculated it can be compared to the equivalent costing single fidelity Kriging model. Figure 26 illustrates only those points of these sampling plans for each function where the multi-fidelity model outperforms the equivalent costing Kriging model in terms of both r^2 and RMSE. Figures 27, 28 and 29 illustrate the same points but collapsed down onto two dimensions. Analyzing these figures and considering only those regions where performance is consistently better for all four test functions a number of important results can be observed.

Figures 27 and 29 illustrate that converting more than 80% of the total evaluation budget into cheap simulations appears to result in poorer performance in two out of the four test functions. Only when predicting the Branin and Trid functions is the multi-fidelity approach better when more that 80% of the total budget is converted to cheap simulations and this is only at relatively high levels of correlation.

Unlike the previous more restricted investigations, Figures 27 and 29 also illustrate that at least 10% of the evaluation budget should always be converted to cheap function evaluations. Both the investigations of the H34 and Trid functions included cases where less that 6% of the budget was converted to cheap evaluations with the assumption of very cheap 15:1 simulations. Even though these cases resulted in more cheap than expensive simulations the resulting multi-fidelity

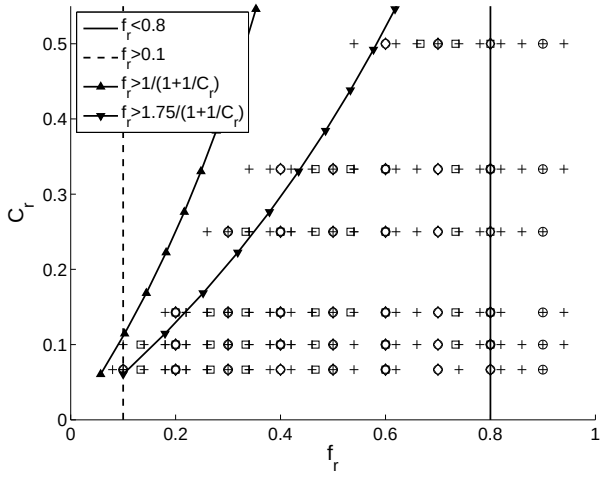


Fig. 27 Overview of settings (ignoring r^2 correlation) for better multi-fidelity than single fidelity performance for all four test functions.

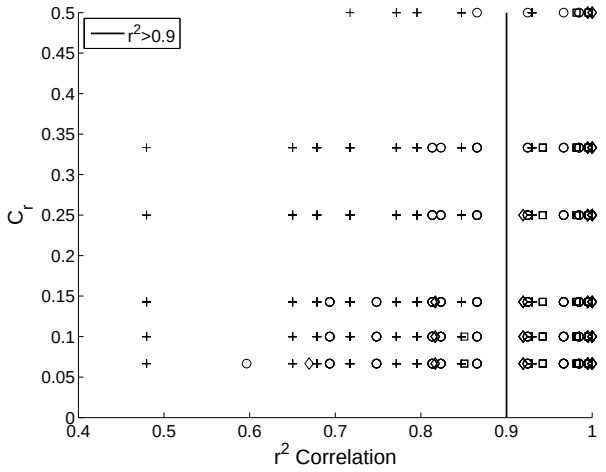


Fig. 28 Overview of settings (ignoring f_r) for better multi-fidelity than single fidelity performance for all four test functions.

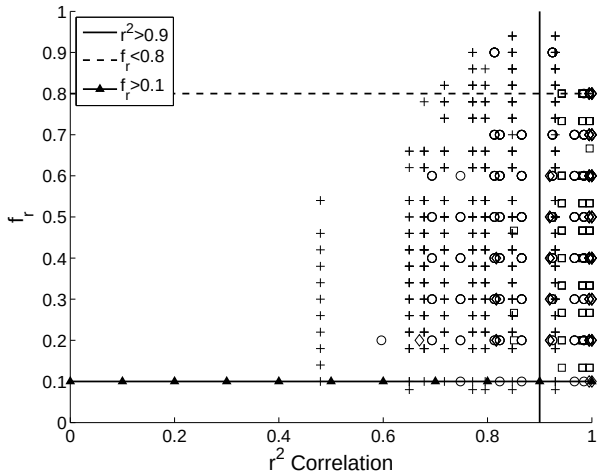


Fig. 29 Overview of settings (ignoring C_r) for better multi-fidelity than single fidelity performance for all four test functions.

model was less accurate than the baseline single fidelity model.

Figure 27 clearly illustrates the importance of having more cheap data in a multi-fidelity model than expensive data. This rule of thumb, defined in the previous investigations, should lead to a case where $f_r > \frac{1}{1+\frac{1}{C_r}}$. However, upon investigating the results illustrated in Figure 27 it is clear that this rule of thumb is not quite conservative enough to meet the requirements of all of the test functions. The inequality of $f > \frac{1.75}{1+\frac{1}{C_r}}$ is a much better fit to those cases where the multi-fidelity model performs better. Both of these constraints have been plotted in Figure 27 where the $f_r > \frac{1.75}{1+\frac{1}{C_r}}$ constraint can clearly be observed to be a much better fit to the performance data especially when a low fidelity evaluation is of a similar cost to a high fidelity evaluation.

Both Figures 28 and 29 clearly demonstrate the impact of the correlation between the low and high fidelity functions on the performance of the multi-fidelity model. Of all of the cases tested there is no point where a multi-fidelity model performs better than a single fidelity model if the r^2 correlation of the underlying function is less than 0.5 and even then with a few exceptions it is only the models of the Trid function which perform consistently better when the r^2 correlation is less than 0.9.

Analyzing the results of this investigation therefore produces four conditions which it could be considered that if fulfilled a multi-fidelity model should out perform an equivalent costing single fidelity model:

1. The correlation between the low and high fidelity function should be reasonably high, $r^2 > 0.9$.
2. No more than 80% of the total evaluation budget should be converted to cheap evaluations, $f_r < 0.8$.
3. More than 10% of the total evaluation budget should be converted to cheap evaluations, $f_r > 0.1$.
4. There should always be slightly more cheap data points than expensive with the inequality, $f_r > \frac{1.75}{1+\frac{1}{C_r}}$, giving a conservative bound for this condition.

If these conditions are not met for an unfamiliar black box function it is recommended that a single fidelity surrogate modeling strategy should be adopted.

8 Engine SFC Optimization

In the previous sections analytical test functions have been employed to ascertain a set of best practice guidelines to help determine when a multi-fidelity surrogate model can be used instead of a single fidelity surrogate

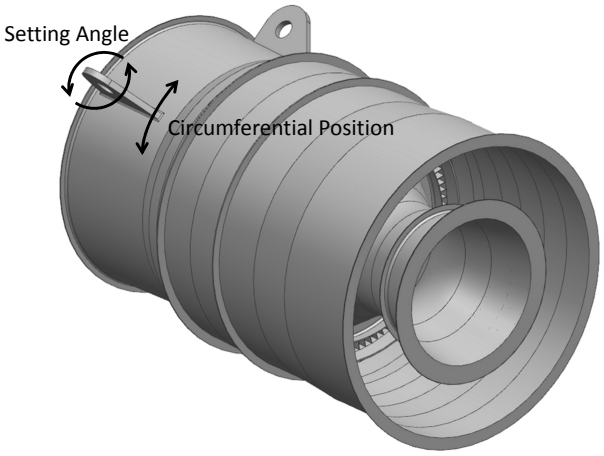


Fig. 30 High pressure compressor casing with modified thrust linkages

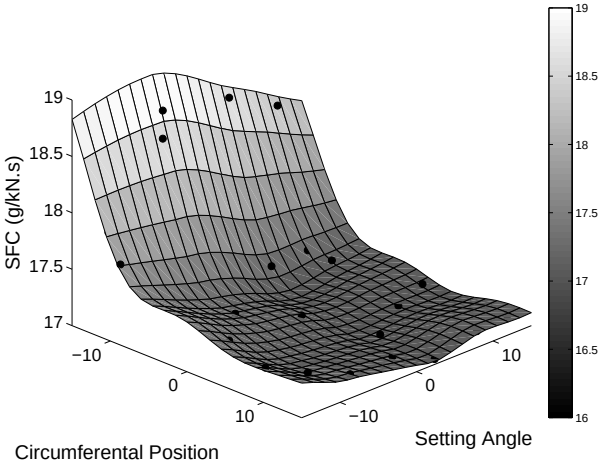


Fig. 31 Variation in SFC with changing thrust link design variables

model. Three engineering design problems from the literature will now be considered to see if they conform to the inferences made using the analytical test functions.

The first of these engineering test problems is a simpler multi-fidelity version of the high pressure compressor optimization taken from Bettebghor et al.[1]. In this optimization the specific fuel consumption (SFC) of the engine is optimized by altering the location of the thrust linkages on the exterior of the high pressure compressor casing, shown in Figure 30. Both the setting angle and circumferential position of these linkages are permitted to vary by ± 15 degrees.

Each engine casing design can be analyzed in two different ways using the propitiatory Rolls-Royce finite element package, SC03. A high fidelity transient thermo-mechanical analysis of the engine can be performed taking approximately 6 days or a low fidelity steady-state mechanical analysis, as used by Bettebghor et al.[1], can be performed in $1/30^{\text{th}}$ the time. In both cases the displacements around the circumference of the casing for each compressor stage are extracted and with a fixed set of rotor platform displacements used to calculate the tip clearance of the compressor. This tip clearance is then used to calculate compressor efficiency and therefore the effect on the SFC of the engine can also be determined.

Figure 31 indicates the 'true' variation in SFC as the thrust linkage setting angle and circumferential position are altered. Given the cost of the high fidelity simulations it is infeasible to perform a full factorial sampling plan of this design space to create an exact representation of the variation in SFC throughout. Instead the surface plot of Figure 31 represents a Kriging model constructed from 30 expensive simulations, indicated by the black dots.

Table 1 Comparison of expensive and cheap engineering simulations used in multi-fidelity examples within the literature

Objective Function	r^2
Specific Fuel Consumption (SFC)	0.972
Compressor Adiabatic Efficiency[2]	0.866
Compressor Massflow Rate[2]	0.849
Compressor Pressure Ratio[2]	0.940
Multi-point Drag Coefficient (M0.75)[24]	0.950
Multi-point Drag Coefficient (M0.2)[24]	1.2×10^{-4}

Even though there is a significant difference in the cost of the finite element simulations used for each fidelity, the results of these simulations are extremely well correlated. Table 1 presents the r^2 correlations between the high and low fidelity models for all of the engineering test problems considered within this paper. Comparing the SFC values resulting from the 30 expensive simulations shown in Figure 31 with their cheap equivalents indicates a relatively high r^2 correlation of 0.972.

Table 2 indicates the accuracy of three surrogate models constructed using three different strategies. Presented in this table is the number of whole engine transient thermo-mechanical simulations (WETTM), the number of whole engine steady-state mechanical (WESM) simulations along with the r^2 correlation, root mean square error (RMSE) and maximum absolute error (MAE) of the surrogate models constructed using these simulations. The three strategies differ in the number of expensive and cheap simulations used in their construction. Cases 1 and 2 use only high fidelity simulations and are therefore Kriging models whereas case 3 uses high and low fidelity simulations and therefore employs

a multi-fidelity model. The sampling plans defining each of the surrogate modeling strategies presented in Table 2 are consistent across all cases. The 5 point sampling plans used in case 2 are optimal subsets of the 10 point sampling plans used in case 1. Likewise the 4 point sampling plans used in case 3 are optimal subsets of the 5 point sampling plans. Case 3 is therefore case 2 with one expensive point replaced by 30 cheap data points of total equivalent cost. A different 10 point 'seed' sampling plan is employed in each of the three tests for each surrogate modeling strategy. The accuracy of the resulting surrogate models are compared to SFC values at the points illustrated in Figure 31.

Given the best practice guidelines defined based on the analytical test functions in this case the multi-fidelity surrogate model would be expected to perform better than a single fidelity model of equivalent cost. The low fidelity simulations are very cheap, the correlation between the models is quite high and only 20% of the total evaluation budget is replaced by cheap simulations. As illustrated by the results presented in Table 2 this is indeed the case. These results indicate that a multi-fidelity model constructed from four expensive simulations and 30 cheap simulations is considerably more accurate than an equivalent costing single fidelity model employing five expensive simulations. The r^2 correlation, RMSE and MAE values are all better for the multi-fidelity model. The accuracy of these models even begins to approach that of a Kriging model constructed from a total of 10 expensive simulations and is therefore twice as expensive.

9 Compressor Rotor Optimization

Brooks et al.[2] compared single and multi-fidelity Kriging in the aerodynamic design optimization of a transonic compressor rotor. The NASA compressor rotor 37[20] was used as the initial design with modifications made to this geometry via 28 design variables controlling blade sweep, lean and skew as well as leading and trailing edge re-cambering at five locations along the blade.

Each design was analyzed using three-dimensional computational fluid dynamics (CFD), in this case employing the HYDRA flow solver[18]. The overall aim of the optimization was to maximize the stage isentropic efficiency of the rotor whilst minimizing the variation in the stage pressure ratio to within 1% and the massflow rate to within 0.5% of those of the baseline rotor.

In order to perform a multi-fidelity optimization cheap data was obtained using a coarse mesh model whilst expensive data was obtained using a fine mesh. A 2.5mm fillet at the hub blade intersection was included in the expensive simulation but absent in the cheap simulation. In this case the cheap model is approximately one third the cost of the expensive model.

Brooks et al.[2] compared the accuracy of the surrogate models created via cross-validation. The multi-fidelity models of the objective function and constraints were observed to be more accurate than the equivalent single fidelity models at the end of the optimization. The final multi-fidelity model of adiabatic efficiency for example had a r^2 of 0.93 and a RMSE of 2.2×10^{-3} while the single fidelity model had a r^2 of 0.67 and a RMSE of 6.5×10^{-3}. The results for the pressure ratio were even better with the multi-fidelity model exhibiting a r^2 of 0.99 and a RMSE of 1.4×10^{-3} compared to the single fidelity model's r^2 of 0.025 and RMSE of 0.164. The massflow rate also showed a considerable improvement with the r^2 rising from 8.6×10^{-5} to 0.96 and the RMSE reducing from 0.29 to 6.4×10^{-4}. Of course not only were the surrogate models more accurate but the final design was also better with a 2.34% improvement in efficiency obtained compared to a 1.79% improvement with standard Kriging.

The application of multi-fidelity Kriging by Brooks et al.[2] was obviously a considerable success but how does the cheap data used in this optimization correlate to the expensive data? Table 1 presents the r^2 correlation between the adiabatic efficiencies, massflow rates and pressure ratios obtained from the cheap and expensive simulations[1]. These results indicate that there is a very good correlation between the design metrics obtained from the cheap and expensive simulations. Comparing this to the results of the analytical test functions it is observed that this is well within the bounds for creating a useful multi-fidelity model observed earlier. The correlations between each of the objectives and constraints are high, although a little short of the previously defined r^2 constraint. Half of the total evaluation budget is replaced by cheap simulations for the initial DoE which means that $f_r = 0.5$ which falls within the $f_r > 0.1$ and $f_r < 0.8$ bounds and given that $C_r = 0.333$, this case also falls within the $f_r > \frac{1.75}{1+\frac{1}{C_r}}$ bound.

10 Multipoint Airfoil Optimization

Consider now the multipoint aerodynamic design optimization of a two-dimensional airfoil section taken from Toal and Keane[24]. In this case 2D CFD simulations,

[1] The sampling plan data for this calculation has been kindly provided by Brooks, Forrester, Keane and Shahpar

Table 2 Comparison of SFC predictions using single & multi-fidelity Kriging models

	Case 1			Case 2			Case 3		
WETTM sims	10			5			4		
WESM sims	0			0			30		
r^2	0.970	0.954	0.935	0.815	0.756	0.029	0.979	0.925	0.846
RMSE	0.138	0.153	0.134	0.253	0.243	0.599	0.093	0.169	0.555
MAE	0.329	0.306	0.232	0.543	0.591	1.521	0.159	0.418	1.407

employing the Fluent flow solver, were used to analyze a 2D airfoil section at three different design points, Mach 0.2, Mach 0.75 and Mach 0.8 at fixed lift coefficients of 1.2, 0.5 and 0.45, respectively. A weighted combination of the drag at these three points was then used as a metric of design quality. The baseline RAE-2822 airfoil section was modified by deforming the computational grid using a freeform deformation approach with ten design variables controlling the vertical displacement of ten control points.

Unlike the compressor rotor optimization the fidelity of the CFD simulation remained constant for both levels of the multi-fidelity surrogate model. Instead the number of design points evaluated was altered. The low fidelity data comprised of the airfoil drag coefficient at only Mach 0.75 while the high fidelity data comprised of the weighted drag coefficient at all three Mach numbers. Low fidelity data could therefore be obtained for one third the cost of the high fidelity data although the actual number of low fidelity simulations was much higher given that the drag at Mach 0.75 must be calculated anyway to determine the weighted drag coefficient.

In this example the multi-fidelity optimizations were observed to offer a much faster rate of convergence than the traditional single fidelity approach obtaining between a 10.4% and 15.7% improvement in the weighted drag coefficient for the equivalent of 60 multipoint simulations whereas the single fidelity approach obtained between 0% and 10.4% improvement for the same simulation cost.

As with the previous examples let us consider the adherence of this real world test case to the guidelines defined using the analytical test cases. Assuming a total budget of 150 high fidelity simulations, all of which are used to construct a baseline Kriging model, one third of these are replaced by low fidelity simulations in an equivalent costing low fidelity model, $f_r = 0.333$. This results in a multi-fidelity model constructed from 100 high fidelity simulations and 250 low fidelity simulations, 100 of which correspond to the high fidelity locations with the remaining 150 low fidelity simulations spread throughout the design space. Given that 50 expensive simulations have been replaced by 250 low fidelity simulations, $C_r = 0.2$. This is perhaps slightly counter intuitive given that a low fidelity simulation is one third the cost of a high fidelity simulation but as a simulation at Mach 0.75 is part of the weighted drag coefficient calculation, the 100 high fidelity evaluations that have been carried out provide 100 additional 'free' low fidelity data points.

As with the previous cases Table 1 shows the r^2 correlation between the airfoil drag at Mach 0.75 and the weighted drag coefficient. The Mach 0.75 drag coefficient is clearly very well correlated with the weighted drag coefficient with $r^2 = 0.95$, this case study is therefore well within our $r^2 > 0.9$ guideline. As $f_r = 0.333$ this case study also meets the upper and lower fixed bounds on f_r and given that $C_r = 0.2$ the case study also meets the $f_r > \frac{1.75}{1+\frac{1}{C_r}}$ bound which states that $f_r > 0.29$. Based on these guidelines the multi-fidelity model should be more accurate than the equivalent costing single fidelity model and indeed it is. The above multi-fidelity model achieves a r^2 of 0.957 and RMSE of 2.04×10^{-3} while the equivalent costing Kriging model employing 150 expensive data points has a r^2 of 0.937 and RMSE of 2.59×10^{-3}.

For the above case the multi-fidelity prediction of the weighted drag coefficient clearly works. However, what would happen if we considered breaking some of the guidelines that have been defined? As the weighted drag coefficient is constructed via a weighted summation of three different drag coefficients the other drag coefficients could feasibly be used to provide the cheap data for the multi-fidelity model instead. The drag coefficient at Mach 0.2, for example, contributes a relatively small amount to the overall objective function. This, in conjunction with its relative distinctness in terms of its flow regime, compared to the other design points, means that the drag at Mach 0.2 is badly correlated to the weighted coefficient, as shown in Table 1. When the drag values of 250 simulations at Mach 0.2 are combined with 100 expensive simulations in a multi-fidelity model, the r^2 drops significantly to 0.221 while the RMSE increases to 4.18×10^{-2}. As expected, whilst the other guidelines are still met, the correlation has fallen significantly below the acceptable level re-

Table 3 Impact of variation in number of expensive data points replaced on weighted drag coefficient surrogate model performance

f_r	DoE (Exp + Chp)	r^2	RMSE
0	150 + 0	0.937	2.59×10^{-3}
0.007	149 + 152	0.893	3.96×10^{-3}
0.033	145 + 160	0.891	3.63×10^{-3}
0.067	140 + 170	0.942	2.76×10^{-3}
0.1	135 + 180	0.950	2.57×10^{-3}
0.13	130 + 190	0.966	1.86×10^{-3}
0.167	125 + 200	0.954	2.47×10^{-3}
0.20	120 + 210	0.941	3.08×10^{-3}
0.333	100 + 250	0.957	2.04×10^{-3}

sulting in a multi-fidelity model which performs poorly compared to the equivalent single fidelity model.

Given the formulation of this test case and the 'free' cheap data from the expensive simulations the constraint to ensure that the number of low fidelity data points is greater than the number of expensive points is never broken. Even if only one of the 150 expensive points is replaced there will still be a total of 152 cheap data points as 149 are calculated as part of the high fidelity evaluations. Instead, let us investigate the applicability of the lower bound on f_r, that is at least 10% of the expensive function evaluations should be replaced with cheap function evaluations. To investigate this let us consider the cases presented in Table 3.

Table 3 presents the accuracy of the weighted drag coefficient prediction for a variety of different fractions of the total simulation budget used to calculate additional cheap sample points. Included in Table 3 is the single fidelity Kriging model with 150 sample points and the multi-fidelity case with 100 expensive and 250 cheap data points. In addition to these, seven other cases are presented where the fraction of expensive points replaced by cheap data points is varied from 0.7% to 20%. In all cases the Mach 0.75 simulations are used to provide low fidelity data. Comparing these additional cases to the baseline single fidelity surrogate model it is clear that once the 10% threshold is reached the accuracy of the resulting multi-fidelity model, in terms of both r^2 and RMSE, is superior to the baseline model thereby supporting the guideline minimum.

11 Conclusions

The construction of a multi-fidelity Kriging model is often, and quite correctly, presented within the surrogate modeling literature as an effective way of improving the accuracy of surrogate models and therefore the performance of any surrogate based optimization employing them. The results of the present article, however, go some way to illustrating that such models are not a panacea for the improvement of any blackbox optimization problem and should in fact be applied with some caution otherwise the surrogate models produced may actually be less accurate than their single fidelity equivalent.

In the current article four analytical test functions and three engineering design problems have been used to investigate a number of the key influences on the performance of a multi-fidelity Kriging model relative to a single fidelity model of equivalent cost. In particular the impact of the correlation between the low and high fidelity functions, the relative expense of the functions, the total evaluation budget and how the total evaluation budget should be divided up between low and high fidelity simulations has been investigated. The results of these individual investigations lead to a more extensive study of the interactions between the most important influences on performance thereby resulting in a set of guidelines to help determine if a multi-fidelity model should be used or not. The guidelines derived using the analytical functions were then assessed with respect to three engineering problems from the literature.

The results of the analytical test function investigations indicated that the correlation between the different function fidelities is extremely important in determining if a multi-fidelity model will be more accurate than a single fidelity model and that a large number of cheap data points is beneficial to the accuracy of the multi-fidelity model but only for closely correlated functions. Similarly, there should always be more lower fidelity data used to construct the model than high fidelity data.

Investigating the impact of the number of expensive function evaluations illustrated that while increasing the amount of expensive data improves the quality of a multi-fidelity model regardless of the correlation between cheap and expensive functions a single fidelity model constructed using the same equivalent budget of simulations will still perform better if the correlation between the functions is poor. The level of correlation between the functions therefore overrides the positive impact of including more expensive data.

Varying how the total evaluation budget is split between the cheap and expensive functions demonstrated that for functions of similar costs the level of correlation plays an even more important role in the accuracy of the final multi-fidelity model. In such cases the evidence suggests that the high level of correlation enables the surrogate to cope with the much smaller amount of high fidelity data available.

A further, more in-depth analysis of the simultaneous impact of function r^2 correlation, relative function cost, C_r, and fraction of expensive simulations replaced with cheap simulations, f_r, helped to defined a set of simple guidelines which can be used to help determine whether a single or multi-fidelity Kriging model should be used:

1. The correlation between the low and high fidelity function should be reasonably high, $r^2 > 0.9$.
2. No more than 80% of the total evaluation budget should be converted to cheap evaluations, $f_r < 0.8$.
3. More than 10% of the total evaluation budget should be converted to cheap evaluations, $f_r > 0.1$.
4. There should always be slightly more cheap data points than expensive with the inequality, $f_r > \frac{1.75}{1+\frac{1}{C_r}}$, giving a conservative bound for this condition.

Successful multi-fidelity surrogate models of specific fuel consumption, compressor rotor performance and multipoint drag performance from the literature were observed to closely follow the above guidelines. Variations in the definition of the multipoint drag case study which led to a model not meeting the above guidelines in terms of minimum r^2 correlation and minimum amount of expensive data replaced by cheap data were also demonstrated to result in a multi-fidelity surrogate model less accurate than the equivalent costing single fidelity model. These results therefore add some weight to the validity of the presented guidelines.

While the above study attempts to be quite thorough there are a number of aspects of multi-fidelity Kriging which could be investigated further. Firstly, the impact of problem complexity was not taken into account in the present study and may play some role in the variation in the results observed for the Trid function compared to the other three analytical functions. The Branin, Paciorek and Hartmann H34 functions are relatively complex in terms of their response, they are multi-modal with a number of local minima, the Trid function however, is convex with a single minima. This may help to explain the ability of the multi-fidelity model to represent the Trid function even when the correlation between the two functions is relatively low. This result suggests that the above guidelines might be somewhat conservative if the underlying shape of the function is relatively simple.

Another interesting result of the above investigations is the demonstration that the scaling parameter of a multi-fidelity Kriging model takes no consideration of the fact that the underlying correlation between the two functions might be poor. The optimization of this hyperparameter therefore tends to give the low fidelity Kriging model more importance than it should. An area of further study might therefore be to investigate alternative formulations for the multi-fidelity prediction which would inherently take the level of correlation between the functions into account thereby preventing the multi-fidelity model becoming worse than a single fidelity model constructed from just the high fidelity data used in the multi-fidelity model.

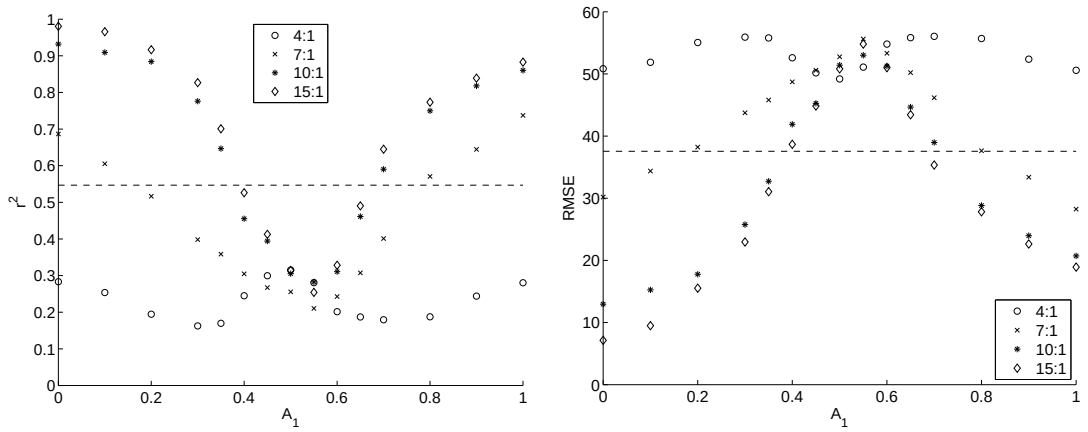
Of course the presented guidelines have been derived with respect to the "true" correlations between the two levels of function fidelity and in a real design optimization this would have to be estimated from a more limited subset. Never-the-less the presented results point towards an effective heuristic which may be embedded within a surrogate modelling toolset in order to automatically select an appropriate single or multi-fidelity approach.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 234344 (www.crescendo-fp7.eu).

References

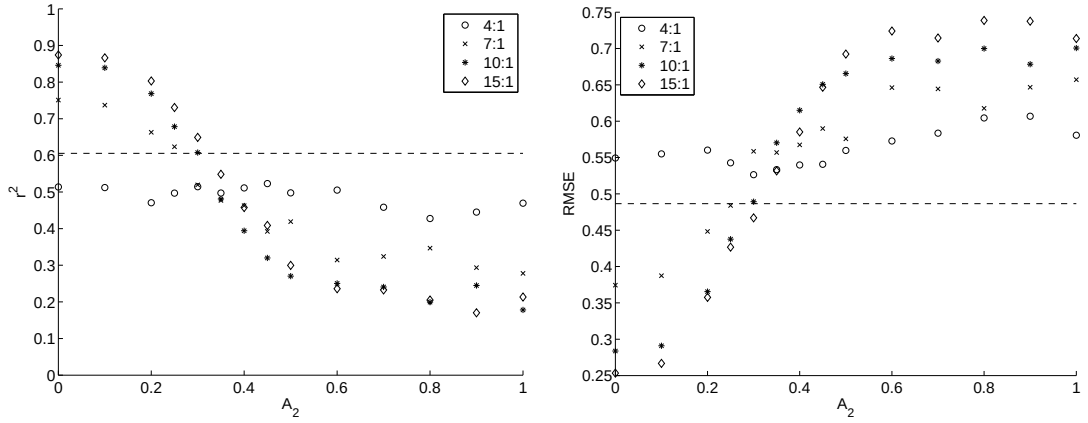
1. Bettebghor, D., Blondeau, C., Toal, D.J., Eres, H.: Bi-objective optimization of pylon-engine-nacelle assembly: Weight vs. tip clearnce criterion. Structural and Multidisciplinary Optimization **48**(3), 637–652 (2013)
2. Brooks, C., Forrester, A., Keane, A., Shahpar, S.: Multi-fidelity design optimisation of a transonic compressor rotor. In: 9th European Turbomachinery Conference, Istanbul, Turkey, 21st-25th March (2011)
3. Forrester, A.: Black-box calibration for complex systems simulation. Phil. Trans. R. Soc. A **368**(1924) (2010). DOI 10.1098/rsta.2010.0051
4. Forrester, A., Keane, A.: Recent advances in surrogate-based optimization. Progress in Aerospace Sciences **45**(1-3), 50–79 (2009). DOI 10.1016/j.paerosci.2008.11.001
5. Forrester, A., Sóbester, A., Keane, A.: Multi-fidelity optimization via surrogate modelling. Proceedings of the Royal Society A **463**(2088), 3251–3269 (2007). DOI 10.1098/rspa.2007.1900
6. Forrester, A., Sóbester, A., Keane, A.: Engineering Design via Surrogate Modelling. Wiley-Blackwell (2008)
7. Ghoreyshi, M., Badcock, K., Woodgate, M.: Accelerating the numerical generation of aerodynamic models for flight simulation. Journal of Aircraft **46**(3), 972–980 (2009). DOI 10.2514/1.39626
8. Goldberg, D.: Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley (1989)
9. Han, Z., Görtz, S.: Hierarchical kriging model for variable-fidelity surrogate modeling. AIAA Journal **50**(9), 1885–1896 (2012)
10. Han, Z., Görtz, S., Zimmermann, R.: Improving variable-fidelity surrogate modeling via gradient-enhanced kriging and a generalised hybrid bridge function. Aerospace Science and Technology **25**(1), 177–189 (2013)
11. Han, Z., Zimmermann, R., Görtz, S.: Alternative cokriging model for variable-fidelity surrogate modeling. AIAA Journal **50**(5), 1205–1210 (2012)

12. Jones, D.: A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization **21**(4), 345–383 (2001). DOI 10.1023/A:1012771025575
13. Jones, D., Schonlau, M., Welch, W.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization **13**(4), 455–492 (1998). DOI 10.1023/A:1008306431147
14. Kennedy, M., O'Hagan, A.: Predicting the output from a complex computer code when fast approximations are available. Biometrika **87**(1), 1–13 (2000). DOI 10.1093/biomet/87.1.1
15. Kirkpatrick, S., Gelatt, C., Vecchi, M.: Optimization by simulated annealing. Science **220**, 671–680 (1983)
16. Krige, D.: A statistical approach to some basic mine valuation problems on the witwatersrand. Journal of the Chemical, Metallurigical and Mining Engineering Society of South Affrica **52**(6), 119–139 (1951). DOI 10.2307/3006914
17. Kuya, Y., Takeda, K., Zhang, X., Forrester, A.: Multifidelity surrogate modeling of experimental and computational aerodynamic data sets. AIAA Journal **49**(2), 289–298 (2011). DOI 10.2514/1.53410
18. Lapworth, B., Shahpar, S.: Design of gas turbine engines using cfd. In: ECCOMAS (2004)
19. Queipo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P.: Surrogate-based analysis and optimization. Progress in Aerospace Sciences **41**, 1–28 (2005). DOI 10.1016/j.paerosci.2005.02.001
20. Reid, L., Moore, R.: Design and Overall Performance of Four Highly-Loaded, High-Speed Inlet Stages for an Advanced, High-Pressure-Ratio Core Compressor. NASA TP-1337 (1978)
21. Sacks, J., Welch, W., Mitchell, T., Wynn, H.: Design and analysis of computer experiments. Statistical Science **4**(4), 409–435 (1989). DOI 10.2307/2245858
22. Simpson, T., Peplinski, J., Kock, P., Allen, J.: Metamodels for computer-based engineering design: Survey and recommendations. Engineering with Computers **17**(2), 129–150 (2001). DOI 10.1007/PL00007198
23. Toal, D., Bressloff, N., Keane, A., Holden, C.: The development of a hybridized particle swarm for kriging hyperparameter tuning. Engineering Optimization **43**(6), 675–699 (2011). DOI 10.1080/0305215X.2010.508524
24. Toal, D., Keane, A.: Efficient multi-point aerodynamic design optimization via co-kriging. Journal of Aircraft **48**(5), 1685–1695 (2011). DOI 10.2514/1.C031342
25. Wankhede, M., Bressloff, N., Keane, A.: Combustor design optimisation using co-kriging of steady and unsteady turbulent combustion. In: Proceedings of ASME Turbo Expo 2011 (2011)
26. Yamazaki, W., Mavriplis, D.: derivative-enhanced variable fidelity surrogate modeling for aerodynamic functions. AIAA Journal **51**, 126–137 (2013)



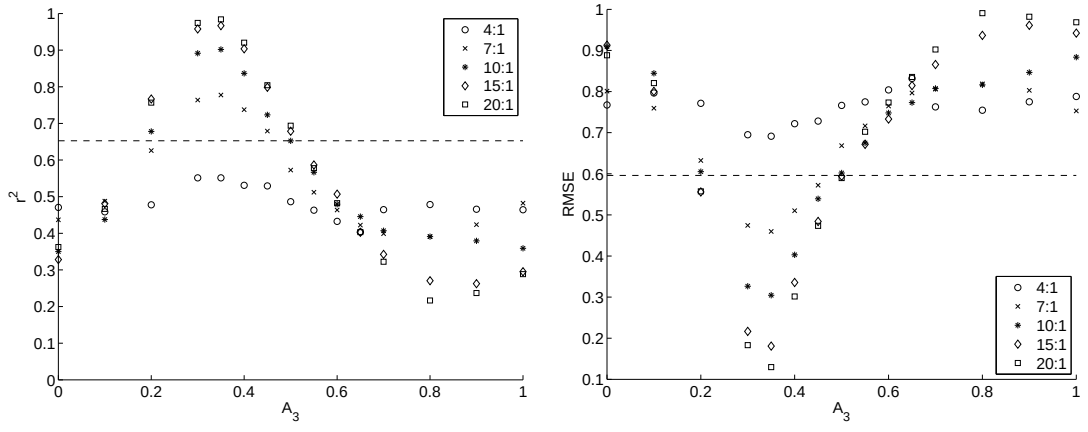
(a) (b)

Fig. 6 Multi-fidelity Kriging prediction r^2 (a) and RMSE (b) of the Branin function with changing cost ratio



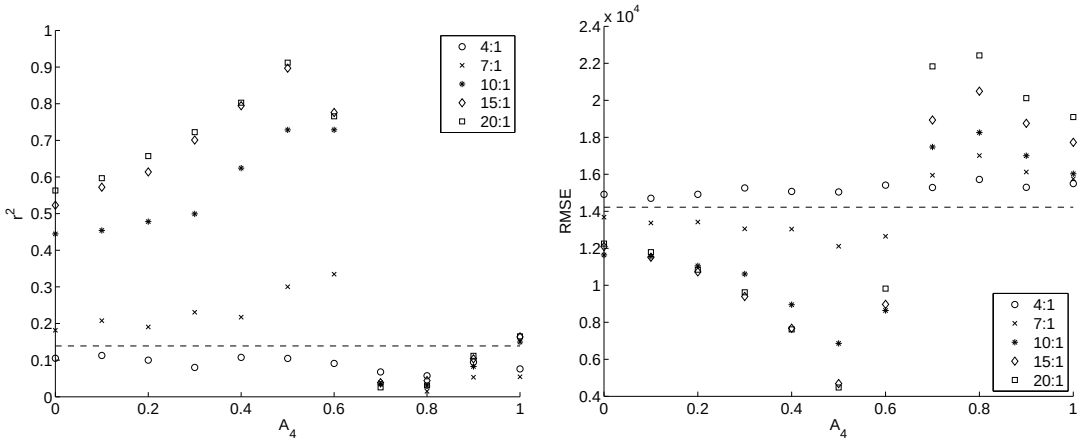
(a) (b)

Fig. 8 Variation in multi-fidelity Kriging prediction of the Paciorek function with changing 'cheap' function cost ratio and A_2



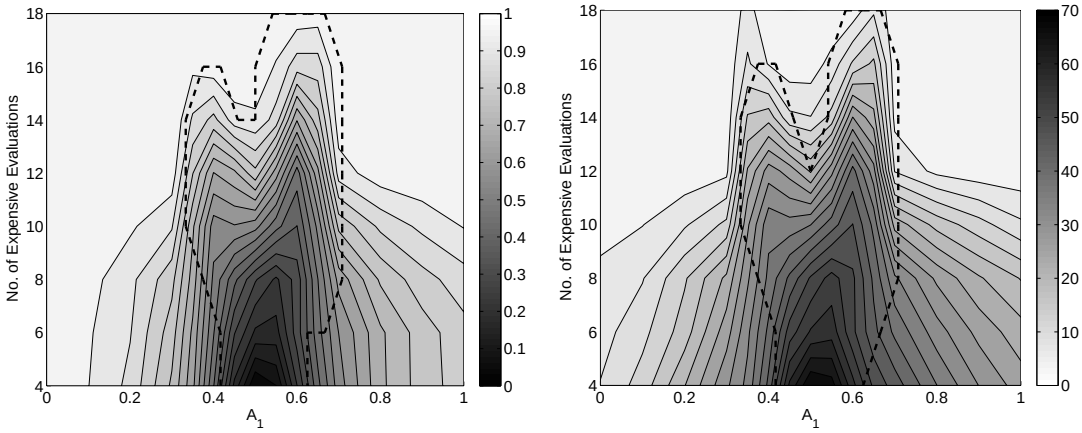
(a) (b)

Fig. 10 Variation in multi-fidelity Kriging prediction of the Hartmann H34 function with changing 'cheap' function cost ratio and A_3



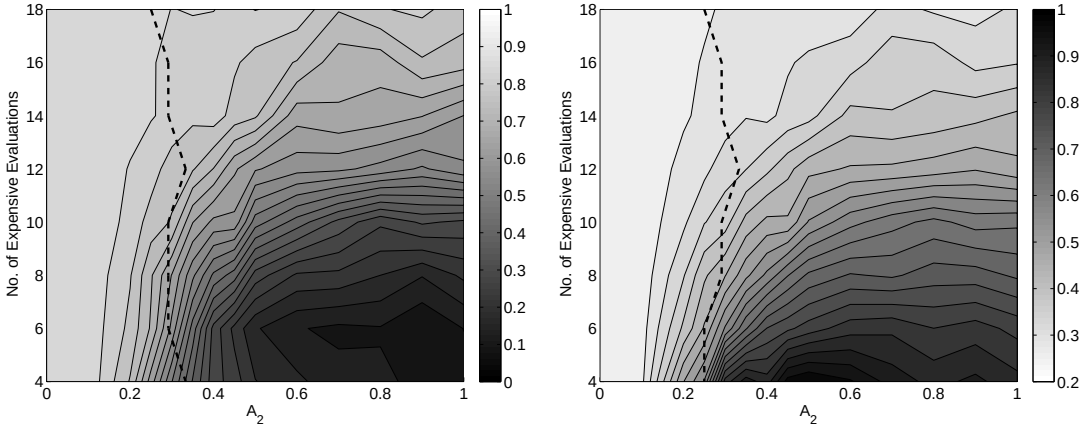
(a) (b)

Fig. 12 Variation in multi-fidelity Kriging prediction of the 10D Trid function with changing 'cheap' function cost ratio and A_4



(a) (b)

Fig. 18 Multi-fidelity prediction r^2 (a) and RMSE (b) of the Branin function with changing no. of expensive simulations, the region outside of the dotted line is where the multi-fidelity model is more accurate than an equivalent cost single fidelity model



(a) (b)

Fig. 19 Multi-fidelity prediction r^2 (a) and RMSE (b) of the Paciorek function with changing no. of expensive simulations, the region to the left of the dotted line is where the multi-fidelity model is more accurate than an equivalent cost Kriging model

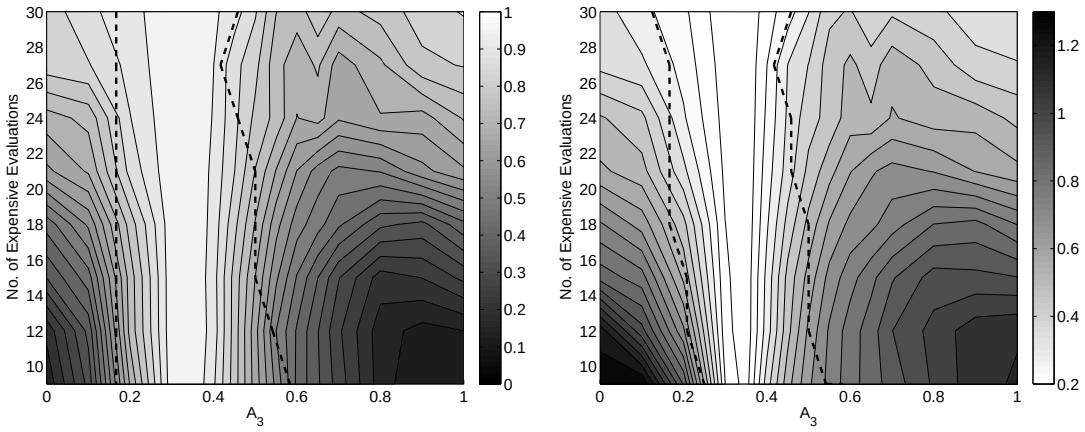


Fig. 20 Multi-fidelity prediction r^2 (a) and RMSE (b) of the Hartmann H34 function with changing no. of expensive simulations, the region bound by the dotted lines is where the multi-fidelity model is more accurate than an equivalent cost Kriging model

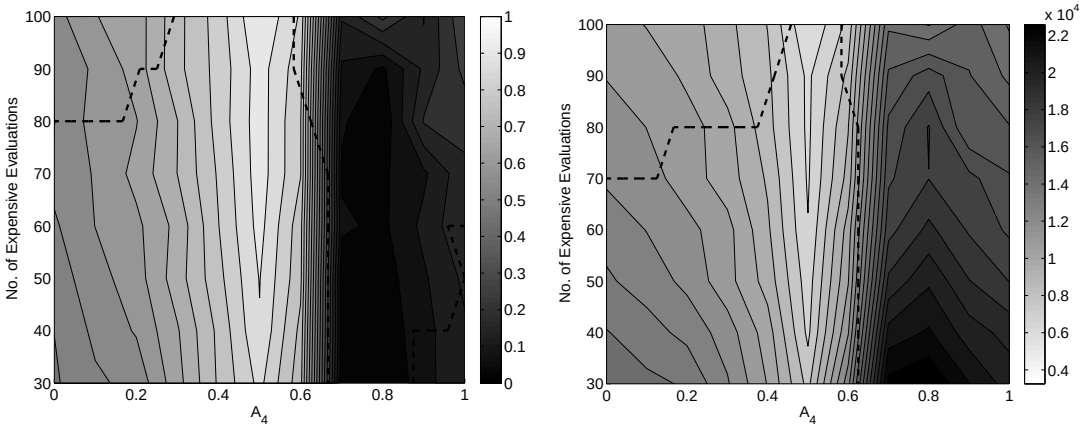


Fig. 21 Multi-fidelity prediction r^2 (a) and RMSE (b) of the Trid function with changing no. of expensive simulations, the region bound by the dotted lines is where the multi-fidelity model is more accurate than an equivalent cost Kriging model

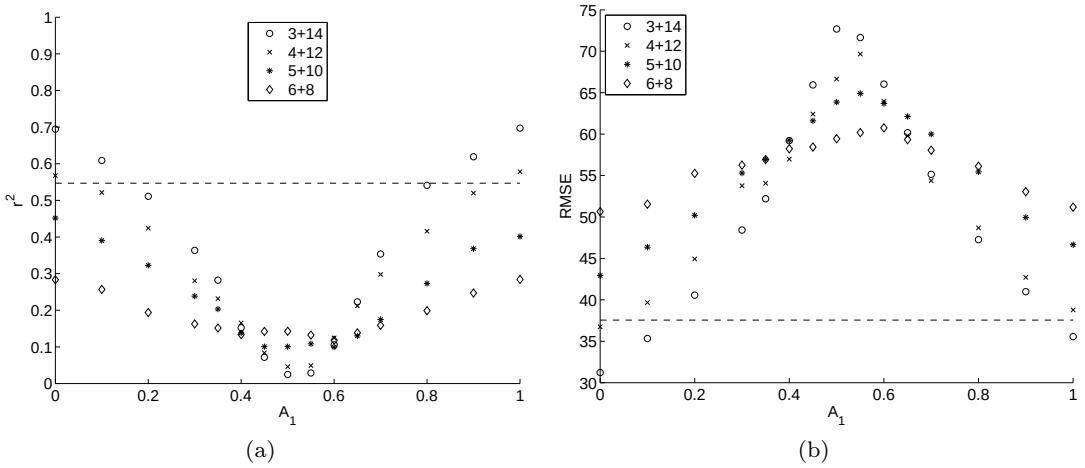
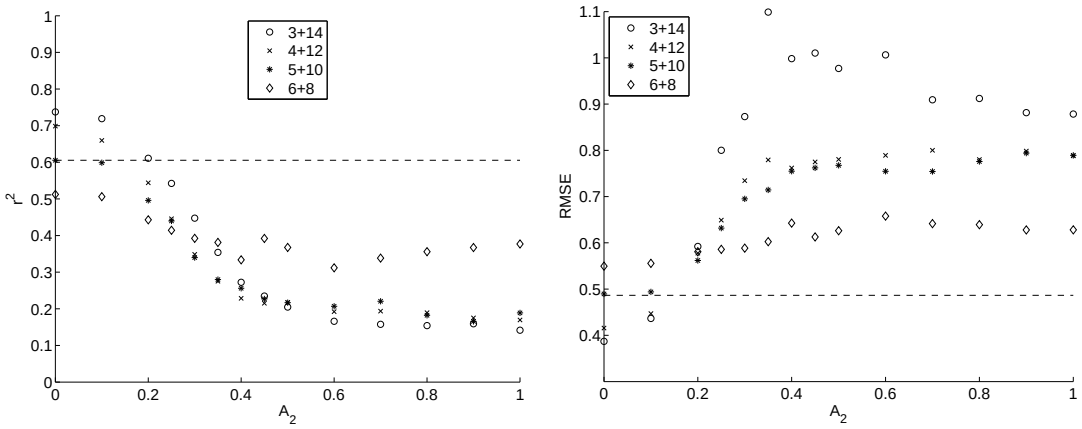
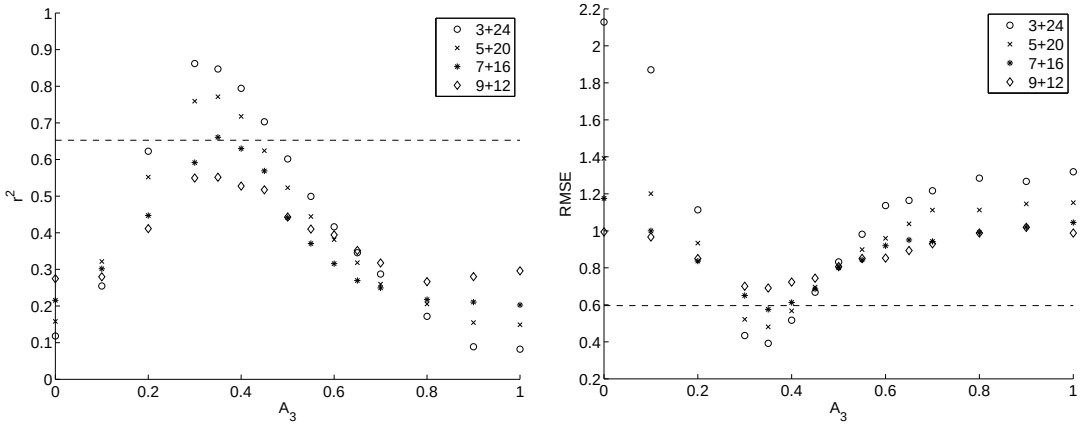


Fig. 22 Multi-fidelity Kriging prediction r^2 (a) and RMSE (b) of the Branin function with changing no. of expensive and cheap simulations for a fixed 2:1 cost ratio



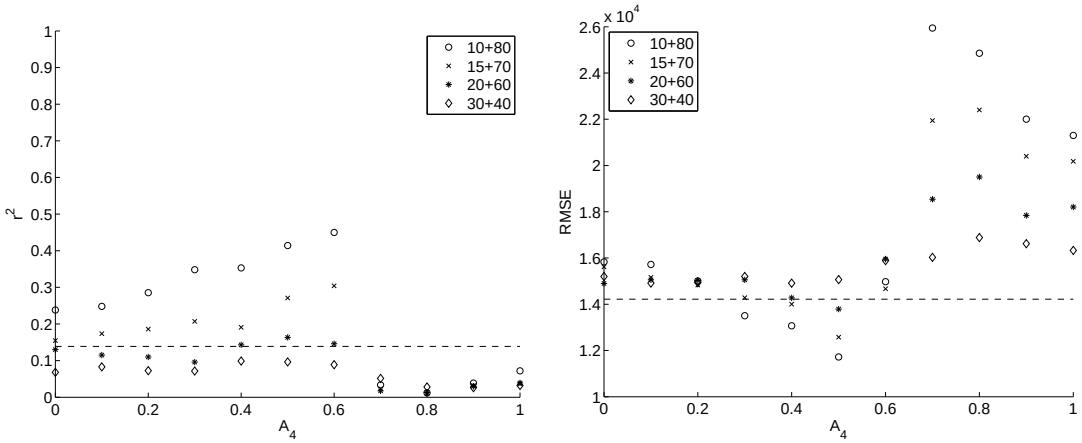
(a) (b)

Fig. 23 Multi-fidelity Kriging prediction r^2 (a) and RMSE (b) of the Paciorek function with changing no. of expensive and cheap simulations for a fixed 2:1 cost ratio



(a) (b)

Fig. 24 Multi-fidelity Kriging prediction r^2 (a) and RMSE (b) of the Hartmann H34 function with changing no. of expensive and cheap simulations for a fixed 2:1 cost ratio



(a) (b)

Fig. 25 Multi-fidelity Kriging prediction r^2 (a) and RMSE (b) of the Trid function with changing no. of expensive and cheap simulations for a fixed 2:1 cost ratio