Structural and Multidisciplinary Optimization (2018) 57:2127–2142
https://doi.org/10.1007/s00158-018-2001-8

RESEARCH PAPER

Analysis of dataset selection for multi-fidelity surrogates for a turbine problem

Zhendong Guo[1] · Liming Song[1] · Chanyoung Park[2] · Jun Li[1] · Raphael T. Haftka[2]

Received: 28 December 2017 / Revised: 20 April 2018 / Accepted: 1 May 2018 / Published online: 25 May 2018

Abstract

Multi-fidelity surrogates (MFS) have become a popular way to combine small number of expensive high-fidelity (HF) samples and many cheap low-fidelity (LF) samples. In some situations LF samples can come from multiple sources and sometimes the HF samples alone can obtain a more accurate surrogate than the combination (HF&LF). Therefore this paper considers using maximum likelihood (ML) and cross validation (CV) to select the dataset leading to best surrogate accuracy, when multiple sample sources are available. The kriging and co-kriging techniques were employed to build surrogates. Unlike conventional model selection, the multi-fidelity datasets selection by ML and CV has to compare the surrogate accuracy of different true functions. The effectiveness of ML and CV is examined through a two-variable turbine problem, where samples can come from one HF and two LF models. The indicators were used to select between using only HF samples or combining them with one set of LF samples or the other. The best selection proved to depend on the design of experiments (DOE), and so datasets were generated for a large number of DOEs. It was found the CV and ML worked relatively well in selection between two LF sample sources for MFS. When selecting between only HF and HF&LF, the ML, which is frequently used in co-kriging hyper-parameter estimation, failed in detecting when the surrogate accuracy of only HF was better than HF & LF. The CV was successful only part of the time. The reasons behind the poor performance are analyzed with the help of a 1D example.

Keywords Multi-fidelity dataset selection · Insufficient data · Kriging/co-kriging · Cross validation · Maximum likelihood · Turbine problem

Nomenclature

R	Correlation matrix
x	Design variable
y	Function values
ρ	Scaling factor between high- and low-function models
σ^2	Variance

Subscript

d	discrepancy
H	High fidelity
L	Low fidelity

Abbreviations

CV	Cross Validation
DF	Discrepancy Function
HF	High Fidelity
LF	Low Fidelity
LHS	Latin Hypercube Sampling
Loo-CV	Leave-one-out Cross Validation
MFS	Multi-Fidelity Surrogate
ML	Maximum Likelihood
RMSE	Root Mean Squared Error
TT	Transient Rotor Blade model with Time Transformation
Transient	Full Transient model

1 Introduction

Surrogate approximations are frequently used in engineering design (Forrester and Keane 2009), in order to reduce computational cost compared to direct simulations such as in computational fluid dynamics (CFD). However, to ensure the accuracy of the surrogate, many samples are required. Usually,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00158-018-2001-8) contains supplementary material, which is available to authorized users.

✉ Raphael T. Haftka
haftka@ufl.edu

[1] Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an 710049, China

[2] Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611-6250, USA

only few samples of direct expensive high-fidelity (HF) simulations are available within limited time budget (Shan and Wang 2010; Liu et al. 2018a). Therefore, one may resort to augmenting a small number of HF simulations with a large number of low-fidelity (LF) but cheap samples to fit the surrogate. This technique is usually referred to as multi-fidelity surrogates (MFS), and has been demonstrated to be an efficient tool in many cases (Forrester et al. 2007; Kennedy and O'Hagan 2000; Fernández-Godino et al. 2016; Park et al. 2017; Liu et al. 2018b).

The LF function is selected beforehand and is fixed in most of the conventional work related to MFS. However, as recently reported in a strength prediction problem, unexpectedly the surrogate built by only 4 or more HF samples was found having better accuracy than an MFS with the same 4 samples, aided by 12 LF simulations in a two-dimensional problem (Zhang et al. 2017). The reason was eventually traced to the mismatch of the signs of small curvature between the HF and LF functions. This highlights the following dilemma: The number and positions of the HF and LF samples are often selected with only minimal information on the shape of the two functions but good information on the cost. However, the utility of combining them does not only depend on the two functions, but also on the lucky placement of the samples. Once samples are available, there is the task of deciding whether an MFS is more or less accurate than the HF surrogate. When samples from multiple LF models are available, there is the additional task of choosing between them.

In conventional model selection, three kinds of indicators are frequently used: (1) Indicators of true accuracy at additional testing points such as the root mean squared error (RMSE) or R^2 (Myers and Montgomery 2002; Martin and Simpson 2005);(2) Indicators using only the given samples and not making any statistical assumptions, such as cross validation (CV) (Viana et al. 2009; Arlot and Alain 2010); (3) Indicators using only the given samples but with assumed probability distributions such as maximum likelihood (ML) (Namura et al. 2017), Akaike information criterion and Bayesian information criterion (Myung and Mark 1997; Neath and Joseph 2012). Provided that plenty of extra samples are available, indicators of type (1) can intuitively measure the lack of fit of the approximation. In this paper we wish to use all the samples for fitting the surrogates, so we use extra points only to measure the effectiveness of indicators of types (2) and (3). Also note that the classic indicators of (2) and (3) were developed for choosing among surrogates that fit the same samples in the same true function space. However, as will be seen in following sections, the true function space is different for samples coming from different information sources. Thereby the objective of this paper is to ask whether CV and ML would still be able to choose between such multi-fidelity datasets.

The investigation was conducted based on the surrogate accuracy of loss function of alternative datasets for a turbine stage. The paper is organized as follows: The surrogates, which are kriging and co-kriging are introduced in Section 2. The classic indicators are discussed in Section 3. Then, by using the flow models introduced in Section 4, dataset selection is carried out for a two-variable turbine problem in Section 5. Then, the typical failures of dataset selection in the turbine problem are illustrated with the help of 1D toy problem in Section 6. Finally, some conclusions are drawn in Section 7.

2 Overview of kriging and co-kriging

2.1 Kriging

Kriging is one of the most popular surrogate techniques. The kriging prediction Y at unknown site $\mathbf{x}$ is built as a trend function $f(\mathbf{x})$ plus a normal random process $Z(\mathbf{x})$ as:

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

where, $f(\mathbf{x})$ is usually a constant, linear or quadratic polynomial, and the constant is most widely used (Namura et al. 2017; Lophaven et al. 2002). $Z(\mathbf{x})$ describes the local features of Y around the n sample points $X = \{\mathbf{x}^{(1)}, \cdots \mathbf{x}^{(n)}\}$, which has zero mean and a covariance function, which usually of the form:

$$\mathrm{cov}\left[Z(\mathbf{x}), Z\left(\mathbf{x}^{(i)}\right)\right] = \sigma^2 \exp\left(-\sum_{h=1}^{k} \theta_h \left\| x_h - x_h^{(i)} \right\|^2\right) \tag{2}$$

where, k denotes the number of variables, σ^2 is the process variance of samples. And θ_h represents the curve roughness hyper-parameter, which determines how quickly the function value changes as $\mathbf{x}$ moves away from a local site $\mathbf{x}^{(i)}$, a high θ_h indicates high frequency function along dimension h. More details can be found in (Lophaven et al. 2002).

2.2 Co-kriging

2.2.1 Co-kriging and estimation of hyper-parameters

Co-kriging is often used to build a surrogate based on data that come from multi-fidelity sources, particularly when only few expensive HF samples are available, while plenty of cheap LF samples can be obtained (Forrester et al. 2007). When the LF model is given, the HF prediction of co-kriging is built as the LF approximation multiplied by a scaling factor ρ plus a Gaussian process $Z_d(\cdot)$ called the discrepancy function (DF), which represents the difference between $\rho y_L(\cdot)$ and $y_H(\cdot)$:

$$\hat{y}_H(\mathbf{x}) = \rho y_L(\mathbf{x}, \boldsymbol{\theta}_L) + Z_d(\rho, \mathbf{x}, \boldsymbol{\theta}_d) \tag{3}$$

where, both $y_L(\cdot)$ and $Z_d(\cdot)$ are usually approximated by kriging, $\boldsymbol{\theta}_L$ and $\boldsymbol{\theta}_d$ are the curve roughness hyper-parameters. When $\rho=0$, co-kriging degenerates to kriging built by only HF samples.

Usually, $y_L(\cdot)$ can be accurately fitted to its original function with sufficient LF samples, so in order to simplify the investigation, we assume there is no approximation error in the LF function. The focus then is on ρ and $Z_d(\cdot)$ (Forrester et al. 2007). Once the LF model is determined, the hyper-parameters of DF including ρ and $\boldsymbol{\theta}_d$ are estimated by maximizing the log-likelihood as:

$$\begin{aligned} f(\rho,\boldsymbol{\theta}_d) &= \max \ln(Ln|\mathbf{d}) \\ &= \max\left\{-\frac{n_H}{2}\ln\left(2\pi\sigma_d^2\right)-\frac{1}{2}\ln\left|R_d(\mathbf{x},\boldsymbol{\theta})\right|\right\} \end{aligned} \quad (4)$$

where, Ln denotes the likelihood function, $\mathbf{d}$ is the vector of DF samples. n_H is the number of HF samples, and $|R_d(\cdot)|$ is the determinant of correlation matrix. In this paper, (4) is used for the multi-fidelity dataset selection process based on ML, as kriging and co-kriging are built based on (4). For explaining the behavior of the ML indicator later on, it is worth noting that the first of the two terms in the right-hand-side of (4) will increase with amplitude of the variability or the range of the fluctuations of the function expressed in σ_d^2. The second term, on the other hand can be shown to increase with the waviness of the function. More details can be found in (Forrester and Keane 2009; Forrester et al. 2007).

For our work, the kriging and co-kriging were implemented with an in-house code (Park et al. 2017). For the tuning of kriging/co-kriging hyper-parameters, the Hooke & Jeeves's pattern search algorithm and numerical techniques including data normalization, Cholesky factorization that are used in DACE toolbox (Lophaven et al. 2002) were employed. Meanwhile, to avoid too bumpy curves, bounds were given to the Gaussian correlation function as $0.6\le\exp\left(-\theta_h\left\|x_h^{(i)}-x_h^{(j)}\right\|^2_{\min}\right)\le 1$, where $\left\|x_h^{(i)}-x_h^{(j)}\right\|^2_{\min}$ denotes the minimum distance between samples along dimension h. In addition, the constant trend function was used for the kriging/co-kriging modeling, as the kriging/co-kriging with constant trend function usually have close and even better accuracy than those by using more complex trend functions (e.g. linear, quadratic and etc.) (Forrester and Keane 2009). Last but most importantly, to obtain the best hyper-parameters of kriging and co-kriging as much as possible, a multi-start strategy was employed in the hyper-parameter tuning process.

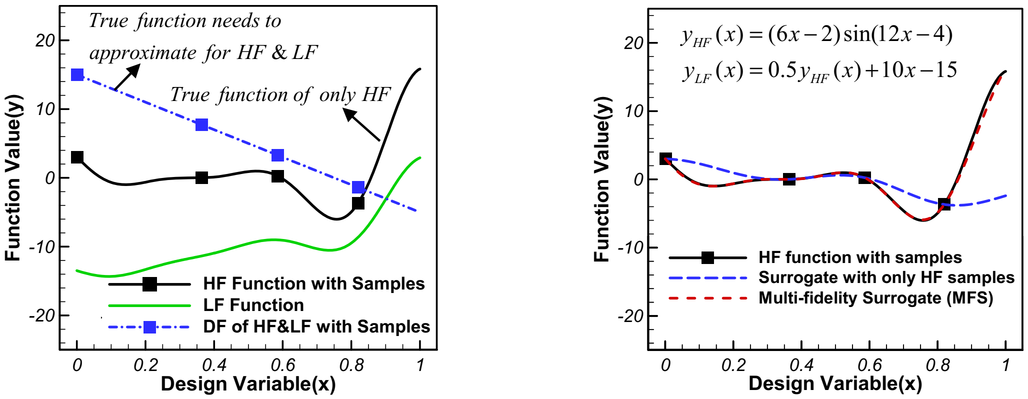
Figure 1 shows a 1D example where both HF and LF samples can be obtained. By solving (4) in combination of HF and LF samples, the best ρ is equal to 2, resulting in a perfectly linear DF, and therefore DF is predicted very accurately with the small number of HF samples (seen in Fig. 1a). Consequently, the multi-fidelity surrogate can perfectly fit the HF function as shown in Fig. 1 (b).

The motivation behind co-kriging is clearly seen as: With the aid of LF models to capture the HF function trend, the original HF curve is modified to be a simpler function (DF) that is likely to be easily fitted by a limited number of HF samples. With different LF models we can expect different DF to fit, so CV and ML need to choose between fitting different true functions, not only different surrogates for a single function.

3 Indicators for multi-fidelity dataset selection

3.1 The challenge of multi-fidelity dataset selection

In conventional model selection, the samples often come from a single source, and the task is to select functional forms (e.g.



(a) Curves of HF, LF and DF functions and samples (b) HF and LF expressions and Surrogates

Fig. 1 Surrogates fit by kriging and co-kriging, HF and LF samples are $x_H=\{0.0007, 0.3638, 0.5862, 0.8198\}$ *and* $x_L=\{0.0007, 0.1548, 0.3638, 0.4934, 0.5862, 0.7148, 0.8198, 0.9932\}$

linear/quadratic and etc.) or surrogates (e.g. kriging/radial basis function and etc.) that can best fit to the true objective function. When multi-fidelity sample sources are available, e.g., both HF and LF sample sources are given; there is the additional task of deciding whether a MFS is more or less accurate than the surrogate built by only HF samples. Further, when multiple LF models (e.g. LF1 and LF2 and etc.) are available, there is one more task of choosing between them to combine with HF samples. We refer to these tasks as dataset selection, which is the focus of this paper.

The classic indicators were tried to tackle with the dataset selection problem, and one challenge is addressed before the selecting process. The classic indicators of cross validation (CV) and maximum likelihood (ML) have been developed for comparing different models with same samples to fit to the same true function (i.e., the situation with samples from a single source). If we want to apply them to compare MFS to a surrogate based on only HF samples, we run into issues because of difference in the number of data. There are probably other ways of addressing this challenge, but here we opt for a simple approach. As plenty of cheap LF samples can be easily obtained, we assume the LF approximation can perfectly fit to its original function; thereby the only source of error comes from the poor fitting of small number of HF samples. So for CV, only the HF samples are perturbed, and for ML, only the likelihood of the discrepancy function (DF) is compared to the ML of the HF alone. This is also consistent with viewing the HF surrogate alone as a special case of the DF for $\rho = 0$.

3.2 Maximum likelihood (ML)

As introduced in Section 2, the ML of DF as (4) is employed by Forrester et al. (2007) in hyper-parameter estimation of co-kriging. During the hyper-parameter estimation by (4), ρvaries from 0 to a certain positive number, where $\rho = 0$ corresponds to the surrogate using only HF samples. It means, the ML of DF has considered the selection between (a) only HF samples ($\rho = 0$) and (b) the combination of HF and LF samples ($\rho > 0$). Actually, the ML of DF can also be used in selection between multiple LF samples sources. Its availability can be derived through the Bayesian posterior probability; more details can be found in Appendix 1. The ML of DF will be called ML in the following sections.

3.3 Cross validation (CV)

The basic idea of CV is: leave out one or several samples, and use the rest of the samples to build the surrogate, and then measure the error at the removed samples. This process is then repeated by leaving out other samples. In the leave-out-one CV (Loo-CV), with m samples this repeats m times and thus makes full use of the samples. In this work, the Loo-CV is used for the dataset selection for multi-fidelity surrogate, and it will not be used to estimate hyper-parameters. Actually, the hyper-parameters of kriging and co-kriging are well estimated beforehand by using (4), and the hyper-parameters are kept constant during dataset selection process by Loo-CV.

The overall accuracy by CV estimation is calculated by the RMSE at CV points as in (5), where n is the number of samples, $\hat{y}_{-i}\left(\mathbf{x}^{(i)}\right)$ denotes the function prediction at $\mathbf{x}^{(i)}$ by the surrogate that is built with $\mathbf{x}^{(i)}$ being removed.

$$CV\_RMSE = \sqrt{\sum_{i=1}^{n} \left(y(\mathbf{x}^{(i)}) - \hat{y}_{-i}(\mathbf{x}^{(i)})\right)^2 / n} \quad (5)$$

4 Flow models and problem description of a turbine stage

Three models are available for the flow analysis of a turbine stage. We assume that these were sampled, and the question is which of the datasets to use for fitting a surrogate. As the main concern of the paper is to examine the performance of CV and ML in dataset selection, so the three models are briefly introduced.

4.1 Design objective of a turbine stage and design variables

The models simulate the flow of a low pressure turbine stage at a flight cruise condition. Figure 2 shows the schematic model of the turbine stage, which consists of one stator vane and one rotor blade. The rotor blade is based on the profile of the popular researched blade, Pak B (Suzen and Huang 2005). A stator vane was designed by referring to E^3 (short for Energy Efficient Engine) turbine stages (Cherry et al. 1982), which can also be replaced by the moving bars for simplicity as the focus is on the flow in the rotor blade (Hodson and Howell 2005). Table 1 shows the specifications of the rotor blade and the stator vane as well, where the left column shows

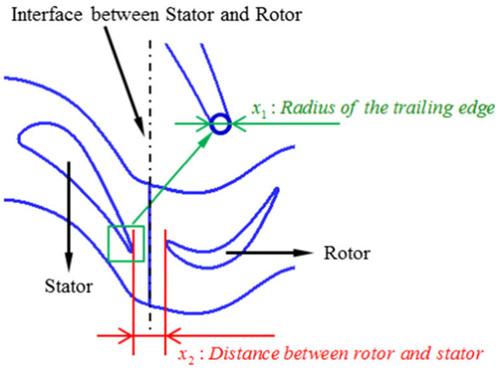
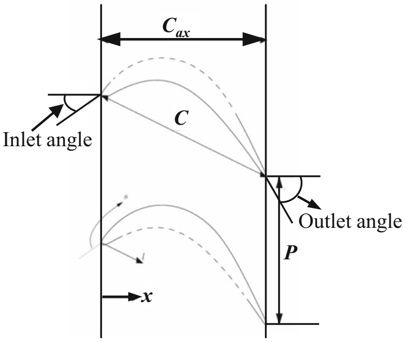

Interface between Stator and Rotor
x_1 : Radius of the trailing edge
Rotor
Stator
x_2 : Distance between rotor and stator

Fig. 2 Turbine geometry and design variables

Table 1 Geometrical specifications of a turbine stage and design variables

Sketch map of Blade	Parameters	Stator	Rotor
 C_{ax} C Inlet angle Outlet angle P x	Blade count	77	77
	Blade height (H)	95.00mm	100.00mm
	Root diameter	350.00mm	350.00mm
	Chord length (C)	48.27mm	33.10mm
	Axial chord length (C_{ax})	34.99mm	30.00mm
	Blade pitch (P)	32.61mm	32.64mm
	Aspect ratio (H/C)	2.91	3.06
	Pitch chord ratio (P/C)	0.68	0.986
	Blade inlet angle	0°	35°
	Blade outlet angle	70.11°	60°
	Rotational speed	0	4500 r/min
Design Variables	x_1	$0.3mm \le x_1 \le 1.0mm$	
	x_2	$5mm \le x_2 \le 15mm$	
Normalized Design Variables	x_1	$0 < x_1 < 1$	
	x_2	$0 < x_2 < 1$	

the definition of the parameters by using rotor blade as an example, the same definition is also suitable for the stator.

The size of the trailing edge of the stator vane (x_1) and the distance between stator and rotor (x_2) will greatly influence the loss and efficiency of the stage (Hodson and Howell 2005). Therefore these two parameters are selected as variables as shown in Fig. 2, and their variable ranges are listed in Table 1. The stage loss, defined as one minus the isentropic efficiency (η_{is}), is chosen as the objective function ((6)), where T^t and P^t denote the total temperature and total pressure, respectively. The subscript 1 and 2 denote the stage inlet and outlet, respectively. γ is the isentropic exponent; more details are available in (Dixon and Cesare 2013).

$$Loss = 1\text{-}\eta_{is} = 1-\left(T_2^t/T_1^t-1\right)/\left(\left(P_2^t/P_1^t\right)^{\frac{(\gamma-1)}{\gamma}}-1\right) \tag{6}$$

4.2 Numerical models and validation

4.2.1 Difference between the three models

Two unsteady Reynolds-averaged Navier-Stokes (RANS) equation solvers including (1) the full transient model (abbreviated as Transient) and (2) the transient rotor blade model with time transformation (abbreviated as TT), and (3) one steady RANS solver (abbreviated as Steady) are used for the performance evaluation (ANSYS CFX-Solver Modeling Guide 2011). The main difference between Steady and Transient is their settings of the interface between stator and rotor (see Fig. 2). The interface technique of frozen rotor is used by the Steady solver, which averages the flow information before it is transported into downstream rotor blade, and hence it cannot consider the transient influence of upstream wake flow on the downstream rotor blade (Hodson and Howell 2005; ANSYS CFX-Solver Modeling Guide 2011). Thereby the loss prediction of Steady solver would be biased. On the contrary, the interface technique of transient rotor stator enables the Transient solver to track the wake blade interactions at periodical instants, so the Transient is expected to have better prediction accuracy than the Steady model.

The difference between TT and Transient is the turbulence model, which plays an important role in simulation accuracy. The Shear Stress Transportation (SST) coupled by $r-\mathrm{Re}_\theta$ transition model is commonly believed to have better accuracy than the SST of fully turbulence model in predicting the separation flow of Pak B (Suzen and Huang 2005). However, the SST coupled by $r-\mathrm{Re}_\theta$ transition is not available in TT; therefore the accuracy of TT would be poorer than that of the Transient. More technical details of Transient, TT and Steady can be found in (ANSYS CFX-Solver Modeling Guide 2011).

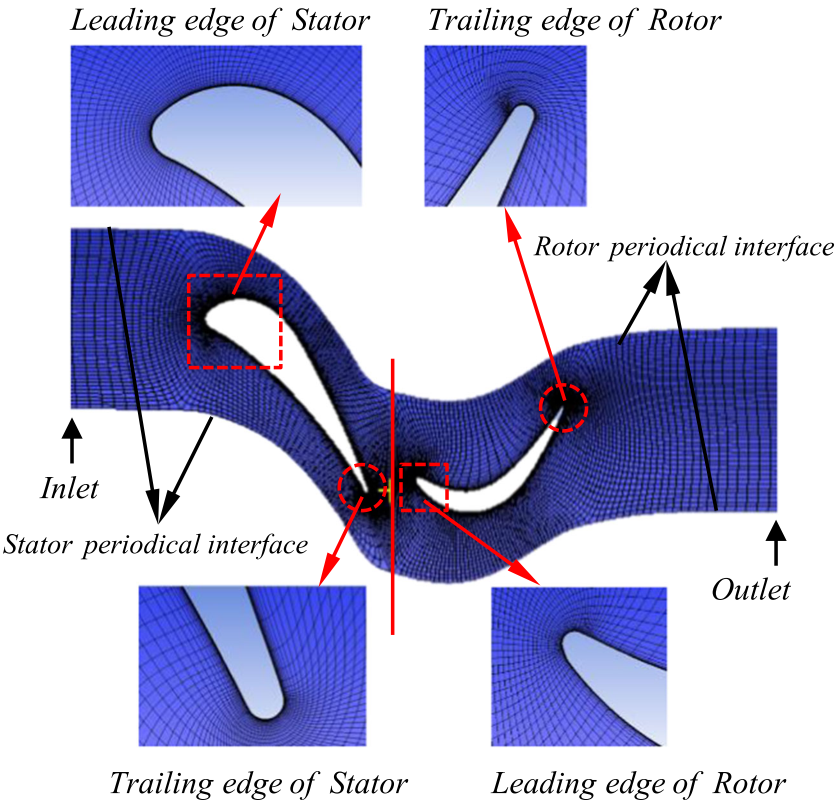
For testing purpose in Section 5, the Transient is regarded as the HF model. Because the TT and Steady have deficiencies in either interface handling between stator/rotor or the turbulence model, they are treated as LF models.

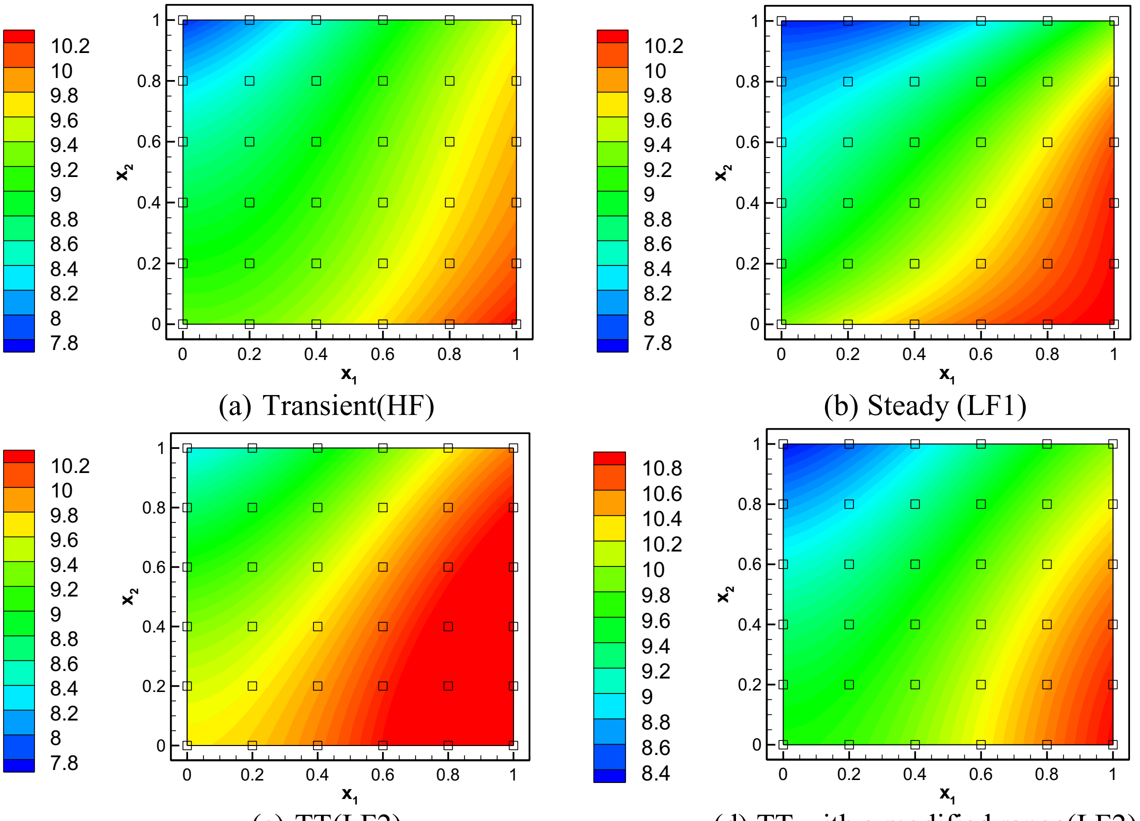
In addition, Table 2 shows the cost of single run of different flow models by using the commercial software CFX 14.5, and each run was implemented on a micro-server by using 8CPUs (2.6GHz). One may argue that the cost of HF model as

Table 2 Computational cost of different flow models

	Computational cost of single run
Transient	20 h
TT	20 h
Steady	1 h

Fig. 3 Computational domain and mesh





(a) Transient(HF)

(b) Steady (LF1)

(c) TT(LF2)

(d) TT with a modified range(LF2)

Fig. 4 The loss contours of the flow models over design space

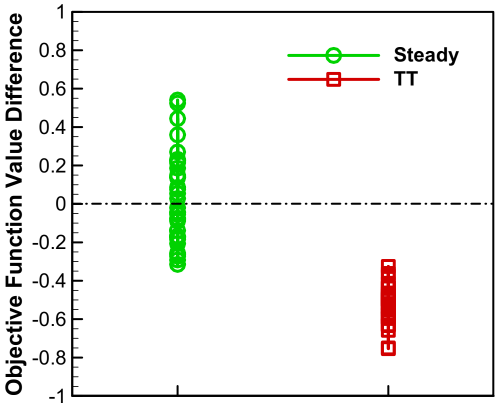


Fig. 5 Distributions of objective function differences between Transient(HF) and Steady (LF1) and that between Transient(HF) and TT (LF2)

Transient and one LF model as TT are the same. However, while it may not be reasonable to use TT for solving an actual design problem, it does provide an additional LF model for testing the ability of CV and ML for choosing the most accurate one.

4.2.2 Mesh parameters and boundary conditions

The settings such as the mesh and boundary conditions are the same for Transient, TT and Steady. Figure 3 shows the mesh grids in the blade to blade view. The H-O-H type grid is employed to ensure grids of high quality. A grid template is adopted to ensure that grids of different designs in space have the same topology, and thus similar grid quality can be guaranteed. To take care of the boundary effects, the first cell to the wall is set to 0.003 mm, correspondingly, the dimensionless distance of the first cell to the wall (ANSYS CFX-Solver Modeling Guide 2011), which is often denoted by y^+, is less than 2.8, i.e., $y^+ < 2.8$. The total number of grid cells is about 760,000.

To simulate the flight cruise condition, the flow conditions were set by referring to E^3 (short for Energy Efficient Engine) turbine stages (Cherry et al. 1982). Specifically, the real gas law is applied. A uniform total pressure and total temperature is imposed at the stator inlet as 106.5 kPa and 873.15 K, respectively. An averaged pressure of 75.86 kPa was set at the rotor outlet. The interface between adjacent stators and rotors are set as rotational periodicity. For the steady flow solver of Steady, the residual error of performance indicators (e.g. the loss in (6)) is set less than 10^{-6} and 300 iteration steps are imposed to ensure convergence. For the transient flow solver as TT and Transient, the same residual error threshold is imposed. Meanwhile, the minimum physical time step is set as 42 in one period that one rotor blade can pass a pitch of the stator blade. 20 periods were calculated to ensure the iteration convergence.

5 Dataset selection for a turbine stage

The multi-fidelity dataset selection was conducted based upon the surrogate accuracy of the turbine stage loss.

5.1 Design space of flow models and accuracy criterion

As described in Section 4, the radius of the stator trailing edge (x_1) and the distance between rotor and stator (x_2) are selected as variables (see Fig. 2 and Table 1). The differences between the loss contours (6) predictions of Transient, TT and Steady are first analyzed. The loss function proved to be somewhat noisy, so for the purpose of this study the response was smoothed by fitting a cubic polynomial to 36 samples from a 6 × 6 grid over the two-dimensional design spaces. One obvious benefit of the data smoothing is that it helps turn this black-box problem to an analytical example, which anybody can try without needing to deal with turbine stage analysis. More details of the cubic regression such as data and regression coefficients are given in Appendix 2. Figure 4 shows the loss contours based on the polynomial fit.

In turbine design, an increment of 0.1% in efficiency (or the decrement of 0.1% loss) is significant (Luo et al. 2015); thereby the accuracy criterion for a reliable approximation is set as the RMSE should be less than 0.1. In Fig. 5, the discrepancy errors of TT and Steady w.r.t. Transient are large, the related RMSE of TT and Steady are 0.2722 and 0.5416, respectively, so they cannot be independently used as alternatives instead of the Transient model.

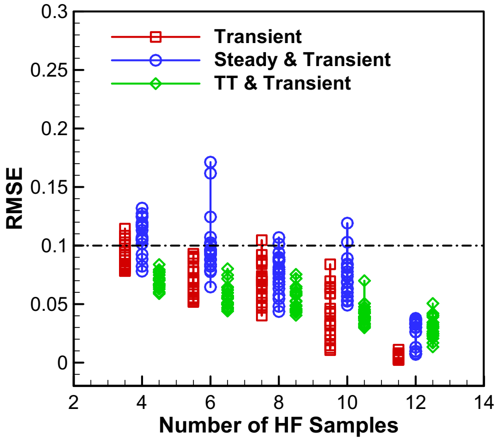


Fig. 6 RMSE Distributions of the surrogate fits to limited number of HF samples. Note that the number HF samples is always even

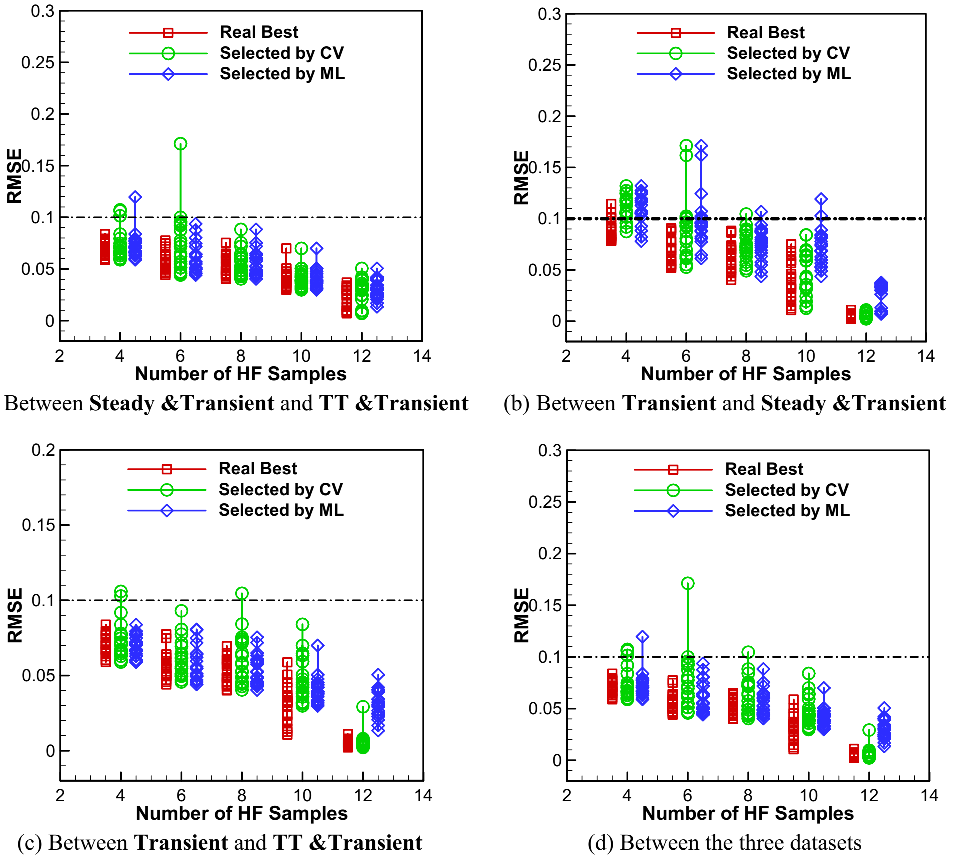
Table 3 The rates of best surrogate accuracy of different datasets

		4HF	6HF	8HF	10HF	12HF
Between the Three	Transient	0/20	3/20	3/20	12/20	19/20
	Steady & Transient	0/20	0/20	3/20	0/20	1/20
	TT &Transient	20/20	17/20	14/20	8/20	0/20
Steady &Transient and TT &Transient	Steady & Transient	0/20	1/20	3/20	0/20	11/20
	TT &Transient	20/20	19/20	17/20	20/20	9/20
TT &Transient and Transient	Transient	0/20	4/20	5/20	12/20	20/20
	TT &Transient	20/20	16/20	15/20	8/20	0/20
Steady &Transient and Transient	Transient	15/20	17/20	14/20	17/20	19/20
	Steady &Transient	5/20	3/20	6/20	3/20	1/20

Figure 4a, b and c compare the three models based on surrogate fit to 36 points using the same scale. However this uniform scale obscures the consistency in trend between Transient and TT, which can be observed with a modified range shown in Fig. 4d. Actually, as seen in Fig. 5, the differences at 36 sites of TT vary in a smaller range ([−0.89,-0.26]) than that of Steady ([−0.44, 0.61]). In addition, the Pearson correlation coefficients between Steady and Transient and that between TT and Transient are 0.9367 and 0.9769, respectively. These confirm that the TT has a better trend consistency than Steady with Transient.

5.2 Settings of kriging and co-kriging models

For MFS, the 36 LF samples will be used, and the number of HF samples will be varied from 4 to 12. Both sets of samples come from the smoothed cubic polynomial. To generate reasonable HF sample distributions as a subset of LF samples, the following criteria were imposed: (i) The HF samples should be evenly distributed over space; (ii) The HF samples should not miss any of the 6 different values of each design variable unless the number of HF samples is smaller than 6; (iii) The distance of each HF sample to its nearest neighbors should be



(a) Between **Steady &Transient** and **TT &Transient**

(b) Between **Transient** and **Steady &Transient**

(c) Between **Transient** and **TT &Transient**

(d) Between the three datasets

Fig. 7 The approximation accuracy of the dataset selected by CV and ML

Table 4 Success rates in selection of the best accuracy by CV

	Steady &Transient and TT &Transient	TT &Transient and Transient	Steady &Transient and Transient	Between the Three
4HF	17/20	17/20	3/20	15/20
6HF	12/20	13/20	6/20	8/20
8HF	16/20	9/20	10/20	8/20
10HF	20/20	6/20	14/20	6/20
12HF	12/20	19/20	18/20	17/20

larger than the minimum distance between LF samples. Appendix 3 provides additional details on HF sampling.

The constant trend function is used for kriging and co-kriging. Data normalizations are conducted based upon the function values of the HF samples. A total of 20 sets of HF samples or DOEs were generated for each prescribed number of HF samples (i.e., 4, 6, 8, 10 and 12) for a total of 100 DOEs.

5.3 Trapezoidal integration-based RMSE for checking performance of CV and ML

The RMSE measures the lack of fit of surrogate by the integration of prediction errors over space. For our case, we collect testing points on a 6 × 6 grid (see Fig. 4) and use trapezoidal integration to calculate the RMSE. This leads to the contributions of each vertex and the edge point to the integral are 1/4 and 1/2, respectively of the contribution of an inner point. This was shown to be a reasonable estimation of RMSE by comparing to the results of a dense grid of 101 × 101 points. More details are given in Appendix 4.

5.4 Selection success and its dependence on the DOE and the number of HF samples

Figure 6 shows the actual error, i.e., the trapezoidal integration-based RMSE at the 6 × 6 grid. Table 3 summarizes the rates of best surrogate accuracy of different datasets. The RMSE of TT& Transient MFS is less than 0.1 for all the sets, even when the 36 TT samples are combined with only 4 HF samples. On the contrary, the RMSEs of Steady & Transient can be higher than 0.1 even with 10 HF samples. This is owing to the better trend consistency between TT and Transient than that between Steady and Transient, though the absolute discrepancy errors between TT and Transient are larger than that between Steady and Transient, as shown in Figs. 4 and 5.

The surrogate of best accuracy is actually DOE-dependent, as shown in Table 3. Though the accuracy of Steady & Transient MFS is usually worse than that of Transient surrogate, it has better accuracy in 18 cases of the 100DoEs shown in Table 3 (see last row of Table 3). With the increase of HF samples, the dataset of best surrogate accuracy changes gradually from TT & Transient to Transient and Steady &Transient as well. For instance, with 12 HF samples, the accuracy of Steady & Transient MFS are shown to be even better than that of TT & Transient for more than half of the cases, in contrast to the situations with 4 to 10 HF samples, where the TT & Transient MFS are often more accurate. So indicators are needed to select dataset of best surrogate accuracy.

Figure 7 compares the accuracy of the datasets selected by CV and ML and that of the real best. Take Fig. 7b for example to illustrate, when selecting between Steady & Transient and Transient with 6 and 8 HF samples in Fig. 7b, the RMSE of the real best meets the accuracy criterion (RMSE<0.1), but the RMSE of the CV- and ML-selected dataset can be much worse. Similar failure situations are also observed in in Fig. 7a, c and d.

The effectiveness of the indicators to select the right datasets is shown in Tables 4 and 5, where the success rate is defined as: Out of the 20 DOEs tested with a given number of HF samples, how many times the datasets of best surrogate accuracy were selected. In 100 sets of DOEs with HF samples changing from 4 to 10, the ML and CV did a relatively good job in selection between the two LF sample sources to combine with HF samples (i.e. Steady & Transient versus TT & Transient). In selection between Transient and Steady & Transient, the performance of ML is poor for all the sets.

Table 5 Success rates in selection of the best accuracy by ML

	Steady &Transient and TT &Transient	TT &Transient and Transient	Steady &Transient and Transient	Between the Three
4HF	19/20	20/20	5/20	19/20
6HF	17/20	16/20	3/20	13/20
8HF	16/20	15/20	7/20	14/20
10HF	20/20	8/20	4/20	8/20
12HF	9/20	0/20	1/20	0/20

Table 6 Summary of success rates in choosing between Transient, Steady &Transient and TT &Transient

		CV	ML
Between the Three	Transient	20/37	0/37
	Steady & Transient	0/4	0/4
	TT & Transient	34/59	54/59
Steady &Transient and TT &Transient	Steady & Transient	6/15	0/15
	TT & Transient	71/85	81/85
TT &Transient and Transient	Transient	22/41	0/41
	TT & Transient	42/59	59/59
Steady &Transient and Transient	Transient	45/82	3/82
	Steady & Transient	6/18	17/18

Further, Table 6 provides a summary of the success rate in identifying the best surrogate for the different selection options when we combine all the sets with different number of samples to obtain 100 DOEs. For example, for the case when all three surrogates are in competition, Transient is best 37 times, Steady &Transient is best 4 times, and TT &Transient is best 59 times. The first row of results in the table shows that CV chose Transient 20 of the 37 times it was best, and ML did not vote for it even once when it was best. It is clear from the table that CV has difficulties in identifying the cases where HF is the best, while ML is a total failure for these cases. When selecting between Steady &Transient and TT &Transient, both indicators showed a clear bias towards TT &Transient. This has only a small effect on the overall rate of success, because in most cases TT and Transient are more accurate. However, out of the 15 cases where Steady &Transient should have been chosen, CV chose it 6 times and ML not even once!

Recalling Section 3.2, the ML in hyper-parameter estimation of co-kriging has considered the selection between only HF samples ($\rho=0$) and HF & LF ($\rho>0$). The above large percentage of failures indicates that, the ML had difficulty in detecting when the HF alone ($\rho=0$) had the best surrogate accuracy (see Table 6). This is unexpected, as the ML is generally believed to be effective in hyper-parameter estimation (Forrester et al. 2007; Kennedy and O'Hagan 2000; Fernández-Godino et al. 2016; Park et al. 2017; Liu et al. 2018b). The performance of CV is also surprising in selection between TT & Transient and Transient. It worked well with small number of HF samples (e.g. 4, 6) but became poor with HF samples in medium size (e.g. 8 and 10). This is opposite to the conventional impression that CV works better with large number of samples. Some insights into the reasons for these failures are provided in the next section with the aid of a 1D toy problem.

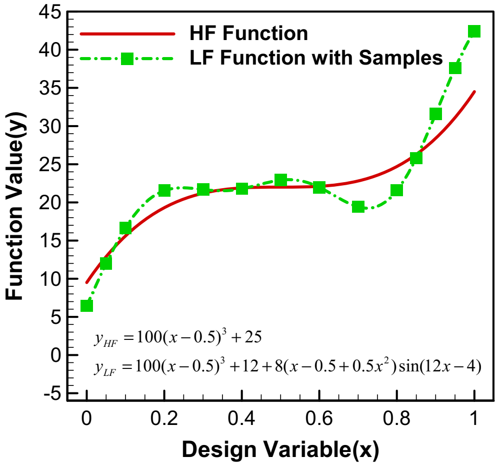


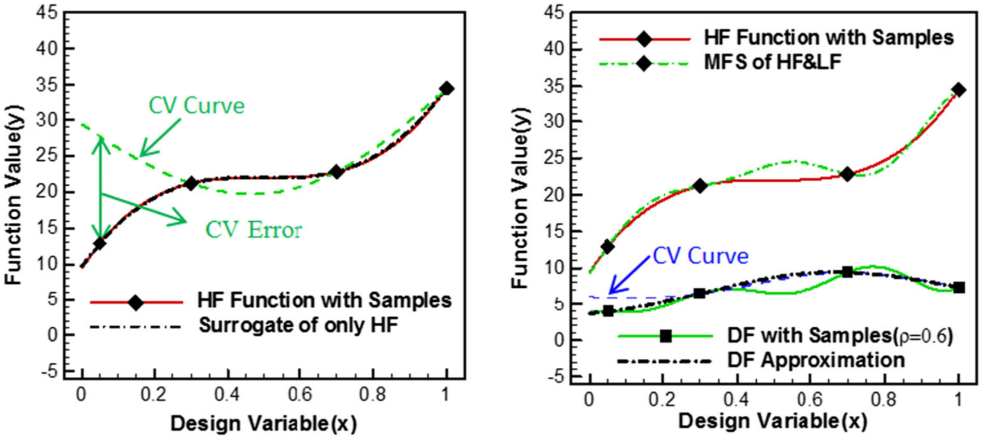
Fig. 8 The HF, LF function of 1D toy problem, the LF samples are at {0.0,0.05,0.1, 0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.85,0.9,0.95,1.0}

6 Analysis of typical selection failures of CV and ML

One-dimensional examples are presented to facilitate understanding of the issues regarding dataset selection of the two-variable turbine problem; the focus is the selection failures between only HF and HF&LF, the characteristics of CV and ML are analyzed. Then, the cause of selection failures of CV and ML in the turbine problem is discussed.

6.1 Multi-fidelity dataset selection in 1D toy problems

Figure 8 shows the HF function and the LF function with samples chosen for exploring the nature of the selection difficulties to choose the HF surrogate when it is more accurate than the MFS. We selected the examples to have two features: (i) the fluctuation range of the discrepancy function is much smaller than the range of fluctuations of the HF function (which is usually the case); (ii) the LF function trend is not in excellent agreement with that of HF function (which happens for the turbine LF models), and thus the discrepancy function can be even bumpier than HF function and more difficult to fit with limited samples. We assume that the LF functions are perfectly fit to their original function with sufficient LF samples. The second feature means that the HF surrogate can often



(a) Only HF (b) HF & LF

Fig. 9 Surrogate of HF alone, HF & LF and example of CV curve, the HF samples are at {0.05,0.3,0.7,1.0}

have better accuracy than the MFS, while the first feature helps mislead the indicators. The dataset selection by CV and ML with 4 and 6 HF samples are discussed separately for CV and ML.

6.1.1 Dataset selection by CV

The failure of CV with small number of samples is not unusual, because then CV tends to overestimate the error estimate, and this is illustrated in Fig. 9. Figure 9a shows that with 4 HF samples the surrogate fits well the function. However, when one sample is removed, the cross validation errors are large. On the other hand, Fig. 9b shows that with the LF function available, the surrogate errors are confined to the discrepancy function, whose fluctuation range is about 3.6 times smaller than the fluctuation range of the HF function, so the cross validation errors are also smaller by about that factor (see Table 7). Consequently, CV fails to recognize that the HF surrogate is more accurate than the MFS.

For the turbine Steady LF this type of failure was observed beginning with small number of samples, because the steady LF trend is not in excellent agreement with transient HF. The TT LF (see Fig. 5) had excellent correlation with the HF; the TT &Transient MFS was more accurate than the HF surrogate for small number of samples, so this problem did not arise much. However, when the number of HF samples increases, the error in the HF surrogate plummeted to the point where this phenomenon appeared. This is because for a two-dimensional function 8 or 10 points is still sparse enough so that cross validation can produce large errors.

When the number of samples becomes sufficient, this difficulty goes away, as shown in Fig. 10 for the CV selection with 6HF samples. Unlike the situation with 4 HF samples, the HF surrogate with leave-one-out samples (CV curves) also fits well the HF function, the CV curves overlapped with the HF surrogate, so they are not presented in Fig. 10a. In contrast, the wavier DF of HF&LF is not fit very well by 5HF samples (when one of the six is left out), and the related CV curve is visually different from the DF function in Fig. 10b. The values of the RMSE errors are shown in Table 7. Similar situation are also observed in Table 4 for the turbine problem, with 12 HF samples, the success rates of CV becomes significantly higher when selecting between Transient and Steady & Transient (or TT & Transient).

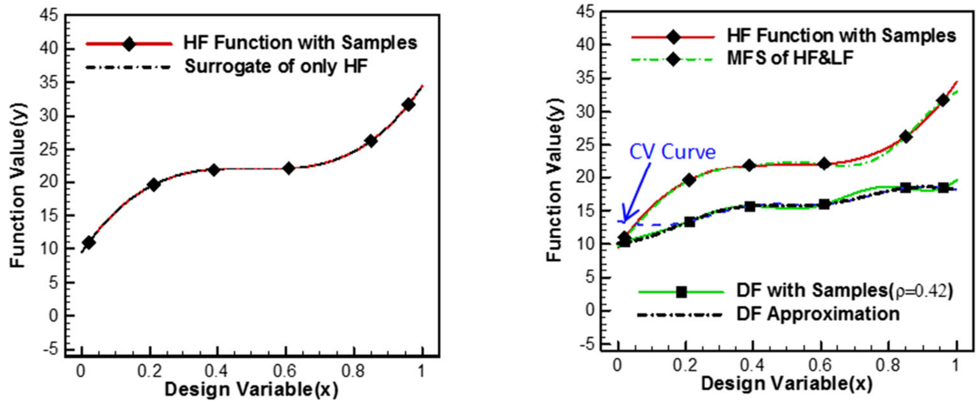
6.1.2 Dataset selection by ML

To understand the ML results, it is worthwhile to recall the discussion of the two components of ML in (4), as $-\ln|R_d(\mathbf{x},\theta)|/2$ and $-n_H\ln\left(2\pi\sigma_d^2\right)\ /2$. The $|R_d(\mathbf{x},\theta)|$ is high

Table 7 RMSE and CV results of only HF and HF & LF MFS

		RMSE	CV_RMSE	CV Errors
4HF	Only HF	0.11	13.50	{-15.00, 7.12, -8.70, 19.21}
	HF & LF	1.03	1.08	{-1.25, 0.19, 1.26, -1.02}
6HF	Only HF	0.000	0.1126	{-0.24, 0.049, -0.029, 0.033, -0.048, 0.11}
	HF & LF	0.43	0.90	{-1.85, 0.038, 0.57, -0.84, 0.60, -0.26}

Fig. 10 Surrogate of HF alone, HF & LF and example of CV curve, the HF samples are at {0.0,0.2,0.4,0.6,0.85,0.95}



(a) HF alone(CV curve overlapped with HF surrogate)

(b) HF &LF

Table 8 RMSE and ML results of HF alone and HF & LF MFS

		ML	$-\ln\lvert R_d(\mathbf{x},\theta)\rvert/2$	$-n_H\ln\left(2\pi\sigma_d^2\right)/2$
4HF	Only HF	−1.81	1.84	−3.65
	HF & LF	0.66	0.40	0.26
6HF	Only HF	−0.66	6.78	−7.44
	HF & LF	0.46	−0.095	0.55

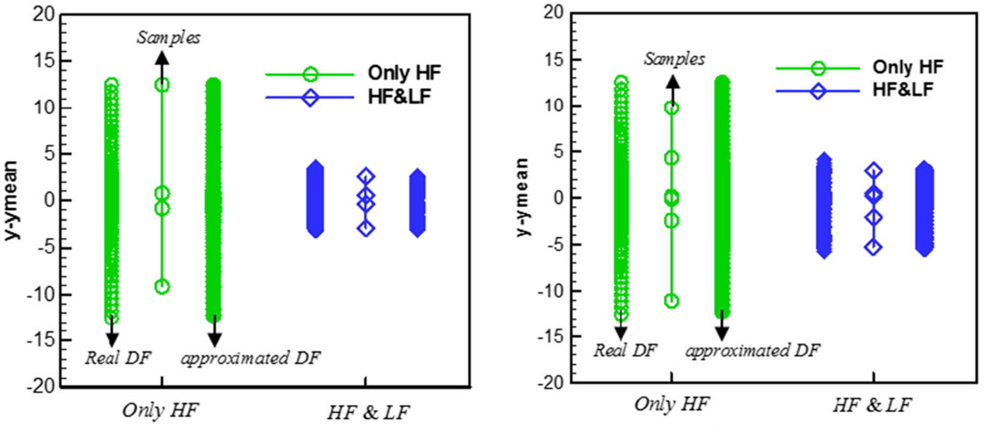
when the function has short wave length, and the σ_d^2 is high when the amplitude of the function fluctuations is high.

Table 8 shows the ML and its components for the toy problem with 4 and 6 HF samples. It is seen that the waviness term $-\ln|R_d(\mathbf{x},\theta)|/2$ is more favorable (more positive) for the less wavy HF. However, the variability term, $-n_H\ln\left(2\pi\sigma_d^2\right)/2$, is more favorable for the MFS and it dominates in both cases. Figure 11 explains these results, showing the fluctuations of DF by subtracting from it the mean function value. For both 4HF and 6HF samples, the fluctuation ranges of DF of HF&LF are much smaller than that of HF alone ($\rho = 0$, its DF is equal to HF function).

This phenomenon of the discrepancy function having a much smaller range of fluctuation than the HF function is not unusual, and so this type of failure of the MF indicator to select the HF cannot be expected to be rare. Recalling Table 6, when selecting between Steady &Transient and Transient, the ML had 3 successes out of 82 total cases when the accuracy of Transient surrogate was better than Steady &Transient MFS.

Note that this analysis also may explain the failure of CV and ML to select Steady &Transient for the turbine problem when this MFS was more accurate than the TT &Transient MFS. As was seen in Fig. 5, the range of differences between

Fig. 11 The fluctuations of DF function



(a) 4HF

(b) 6HF

Steady and Transient was much larger than the range of differences between TT and Transient. So the range of the fluctuations of DF for Steady& Transient must be higher than for the DF of TT& Transient. On the one hand, this is responsible for the superior performance of TT& Transient for most cases. However, this larger range misleads the two indicators for the few cases when Steady & Transient was more accurate. In addition, another suggested version of CV, the CV-based pseudo-likelihood (Rasmussen and Williams 2006), has also been tried, but it has similar poor performance as ML, so the related results are not discussed.

7 Conclusions

The performance of cross validation (CV) and maximum likelihood (ML) in multi-fidelity dataset selection was examined through a two-variable problem of a turbine stage. For that problem the indicators were tasked with selecting between a high-fidelity (HF) dataset for constructing a surrogate based only on this dataset, or combining it with one of two low-fidelity (LF) datasets, LF1 and LF2, for constructing multi-fidelity surrogate (MFS). For the MFS, the main fitting challenge is to fit a discrepancy function (DF) to the difference between the HF function and an optimally scaled LF function. Because the DF is different for LF1 and LF2, CV and ML have to compare the accuracy of approximations that fit different true functions. This is different from conventional model selection that considers alternatives for fitting the same true function.

Tests were conducted for 100 DOEs with the number of HF samples ranging from 4 to 12. The most accurate surrogate was found to depend on the DOE. The percentage of DOEs for which the HF surrogate was most accurate increased, as expected, with the number of samples. The MFS using LF2 was more accurate than the MFS using LF1 for most of the DOEs, but not all. It was found the CV and ML did a relatively good job in selection between LF1 and LF2 when LF2 was most accurate (the majority of the time), but they have substantial trouble finding when LF1 led to the best accurate MFS. Similarly, the indicators were relatively successful finding when the MFS was more accurate than the HF surrogate. However, CV often had trouble selecting HF when its surrogate was most accurate, and ML never identified such cases.

To understand the selection failures in the turbine problem we constructed a 1D toy problem that shared with the turbine problem the property that the LF function trend was not in excellent agreement with the HF function. The analysis of the 1D problem established that ML is biased towards selecting the function with the smaller fluctuation range. This biases it in favor of the DF that has a smaller fluctuation range than the HF function, even if the HF surrogate is more accurate. CV can have a similar problem with small number of samples. Similarly, LF2 had a smaller range of DF fluctuations than LF1, which made its MFS more accurate most of the time. However, the smaller range appears to have biased the indicators to select it even when it was not the most accurate. The results raise concerns that may justify further research into the problem of selecting the dataset that will yield the most accurate surrogate.

Acknowledgments The authors would like to express appreciation for the financial support from the China Scholarship Council (201606280218) and the National Natural Science Foundation of China (51676149), for this work under-taken as part of the first author's Ph.D. project. The authors would also like to gratefully acknowledge the use of facility in Structural& Multidisciplinary Optimization Group of University of Florida (http://www2.mae.ufl.edu/mdo/) in carrying out this work.

Appendix 1: ML criterion for choosing between datasets

The ML of DF in Forrester's version (Forrester et al. 2007) is the case when the LF function is given. Actually, it can also be used in selection of samples from alternative LF sources, when fitting a surrogate to the limited HF samples. For such case, the LF sources is treated as another hyper-parameter as D_L, i.e., we have to choose D_L in addition to ρ and θ_d in fitting the co-kriging by (4). The corresponding Bayesian posterior probability is formulated as:

$$P(D_L, \rho, \theta_d|\mathbf{y}_H) \\ = Likelihood\left(\mathbf{y}_H|\tilde{D}_L, \tilde{\rho}, \tilde{\theta}_d\right) \cdot Prior(D_L, \rho, \theta_d)/m(\mathbf{y}_H) \tag{A.1}$$

where, $m(y_H)$ is the marginal distribution of the dataset $\mathbf{y}_H$, $Prior(D_L, \rho, \theta_d)$ is the prior probability of the hyper-parameters D_L, $\rho$and θ_d, the $Likelihood\left(\mathbf{y}_{HF}|\tilde{D}_L, \tilde{\rho}, \tilde{\theta}_d\right)$ is the likelihood of the co-kriging when $\tilde{D}_L, \tilde{\rho}, \tilde{\theta}_d$ are given, which is actually equal to the formulation of (4). The term $m(y_H)$ is a constant w.r.t. the variation of hyper-parameters D_L, ρ, θ_d, thus for the purpose of selection with fixed HF samples, $m(y_H)$ can be discarded. Meanwhile, we do not have prior knowledge of the hyper-parameters, so it is defensible to

Table 9 Data smoothing by using cubic polynomial regressions

	Transient	TT	Steady
Function range	[7.90, 10.23]	[8.40,10.86]	[7.85,10.27]
Adjusted R^2	0.9606	0.9870	0.9585
RMSE	0.0930	0.0578	0.1236
Mean Error	0.0765	0.0428	0.0939
σ	0.1095	0.0678	0.1414
Relative error	4.70%	2.76%	5.84%

Table 10 Regression coefficients of the cubic polynomials of turbine stage efficiency (100%-Loss) for different flow models with normalized design variables

	constant	x_1	x_2	x_1^2	x_1x_2	x_2^2	x_1^3	$x_1^2x_2$	$x_1x_2^2$	x_2^3
Transient	89.7690	2.9245	0.1566	−6.8921	0.1197	−0.0119	2.8428	0.0633	−0.0148	0.0008
TT	90.5742	0.0250	−0.0640	−1.3642	−0.0236	0.0115	−0.0867	0.0536	−0.0045	0.0000
Steady	88.9969	−2.1275	0.7212	1.4796	−0.0636	−0.0620	0.6834	−0.3760	0.0249	0.0017

simplify the justification by using the non-informative prior as $Prior(D_L, \rho, \theta_d) = 1$ (Neath and Joseph 2012). Hence, the ML of DF in (4) may be still useful for the selection between LF samples coming from alternative LF sources.

Appendix 2: Polynomial smoothing of turbine data

As kriging and co-kriging are sensitive to the data noise, which will also influence the selection of CV and ML, and hence make complicate the problem. Therefore, polynomial regression is employed to smooth the data sets. The RMSE, adjusted R^2 (see B.1), and the mean absolute error and the standard deviation (denoted by σ) of polynomial regression (Myers and Montgomery 2002) as well are calculated to inspect the goodness of the dataset.

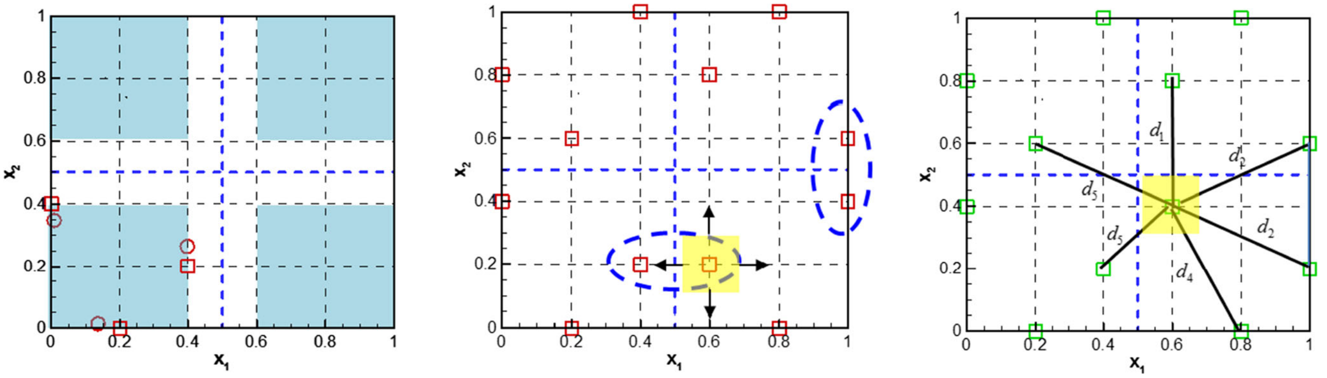
$$adjusted\ R^2 = 1 - \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-p) \right\} / \left\{ \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 / (n-1) \right\} \tag{B.1}$$

where, n is the number of samples, p is the number of polynomial coefficients, y_i and $\hat{y}_i$ are the true and estimated function value of the ith sample, respectively. $\bar{y}_i$ is the mean function value of the samples. In addition, the relative errors of polynomial regressions are also calculated, as dividing the σ by related function range. Table 9 shows the results of cubic polynomials, and Table 10 provides the regression coefficients of cubic polynomials for different flow models. Obviously, the cubic regression can well predict the function trend of different flow models, as the data noises is small.

Appendix 3: HF Sampling strategy

The HF sampling strategy was devised to prevent poor design of experiments for the turbine problem. It is not intended to serve as a general approach, as it is tailored to the specifics of this 2D problem, where samples were available on a grid. The HF sampling is based on the strategy of nearest neighbor sampling (NNS) (Park et al. 2017). The basic idea of NNS is shown in the lower left of Fig. 12a: First, m HF samples are generated independently by using Latin hypercube sampling (abbreviated as LHS). Second, each LHS sample (circles) is moved to its nearest LF site (squares). When the number of HF samples is smaller or equal to the number of LF values in each dimension (e.g. 4 or 6 HF samples), the NNS strategy is directly used.

When the number of HF samples is larger, they are generated as follows: First, the HF samples are sequentially generated by NNS in the four shaded subspaces of Fig. 12; Second, some local samples may not meet the criterion of $d > 0.2$ (seen in Fig. 12b), the violated samples will be moved to its adjacent LF sample location shown by arrows. The objective is to maximize the distance to its neighboring HF samples as $\max\{d_1 + d_2 + \cdots\}$ will be imposed to optimize the sample locations. Similar fine tuning strategy is also implemented in



(a) Generation based on NNS (b) fine tuning (c) Final HF samples

Fig. 12 HF sampling strategy in case of large number of HF samples

Table 11 Averaged RMSE and relative errors of Trapezoidal integration-based RMSE w.r.t. that of 101 × 101 points

		4HF	6HF	8HF	10HF	12HF
All Three	101	0.1098	0.06714	0.05940	0.04349	0.01634
	Trapezoidal	0.1167	0.07555	0.06631	0.05167	0.02020
	Relative Error	6.28%	12.53%	11.63%	18.81%	23.62%
Transient	101 × 101	0.1107	0.06581	0.05912	0.04373	0.01562
	Trapezoidal	0.1179	0.07413	0.06596	0.05184	0.01939
	Relative Error	6.50%	12.64%	11.57%	18.54%	24.14%
Steady & Transient	101	0.1108	0.06691	0.05967	0.04425	0.01604
	Trapezoidal	0.1180	0.07531	0.06660	0.05256	0.01994
	Relative Error	6.50%	12.55%	11.61%	18.78%	24.31%
TT & Transient	101	0.1103	0.06667	0.05976	0.04400	0.01654
	Trapezoidal	0.1175	0.07503	0.06662	0.05228	0.02037
	Relative Error	6.53%	12.54%	11.48%	18.82%	23.16%

Table 12 The changing rate of accuracy order by using the Trapezoidal integration-based RMSE w.r.t that of 101 × 101 points

	Steady &Transient and TT &Transient	TT &Transient and Transient	Steady &Transient and Transient	Between the Three
4HF	1/20	0/20	1/20	1/20
6HF	0/20	2/20	3/20	3/20
8HF	0/20	2/20	3/20	0/20
10HF	1/20	0/20	1/20	0/20
12HF	3/20	0/20	0/20	0/20

the case of small number of HF samples when the distance criterion is violated.

When the number of samples is a multiple of 4, e.g.8, the samples can be evenly distributed in the four subspaces. When the number of samples is not a multiple of 4, e.g. 10, we should have 2 in two subspaces and 4 in other two. The specific number in each subspace is determined randomly.

Appendix 4: Accuracy of trapezoidal integration

Tables 11 and 12 shows the comparison results of the Trapezoidal integration-based RMSE. It is calculated by a dense testing grid of 101 × 101points. The latter can be regarded as accurate enough RMSE owing to sufficient testing samples. Table 11 shows that, the RMSE values estimated by Trapezoidal integration are reasonably close to those of 101 × 101 points. Further, Table 12 shows the changing rate of accuracy order by using Trapezoidal integration-based RMSE; clearly the accuracy order estimated by Trapezoidal integration-based RMSE are in well consistent with that of 101 × 101 points, in other words, The Trapezoidal Integration-based RMSE is accurate enough to judge the selection success of CV and ML in multi-fidelity dataset selection.

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

References

ANSYS, 2010, ANSYS CFX-Solver Theory Guide, Release 13.0. ANSYS Inc., Canonsburg, PA

Arlot S, Alain C (2010) A survey of cross validation procedures for model selection. Stat Surv 4:40–79

Cherry DG, Gay CH, Lenahan DT (1982) Energy efficient engine. Low pressure turbine test hardware detailed design report. NASA CR167956

Dixon, SL, Cesare H (2013) Fluid mechanics and thermodynamics of turbomachinery. Elsevier Inc, Butterworth-Heinemann

Fernández-Godino MG, Park C, Kim NH, Haftka RT (2016) Review of multi-fidelity models. arXIV preprint arXiv:1609.07196. http://arxiv.org/abs/1609.07196

Forrester AIJ, Keane AJ (2009) Recent advances in surrogate-based optimization. Prog Aerosp Sci 45(1–3):50–79

Forrester AIJ, Alexander IJ, Sóbester A, Keane AJ (2007) Multi-fidelity optimization via surrogate modeling. Proc R Soc Lond A Math Phys Eng Sci 463(2088):3251–3269

Hodson HP and Howell RJ. The role of transition in high-lift low-pressure turbines for aeroengines. Prog Aerosp Sci, Vo. 41, No. 6, 2005, pp. 419–454

Kennedy MC, O'Hagan A (2000) Predicting the output from a complex computer code when fast approximations are available. Biometrika 87(1):1–13

Liu HT, Ong YS, Cai J (2018a) A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design. Struct Multidiscip Optim 57(1):393–416

Liu HT, Ong YS, Cai J, Wang Y (2018b) Cope with diverse data structures in multi-fidelity modeling: a Gaussian process method. Eng Appl Artif Intell 67:211–225

Lophaven SN, Nielsen HB and Sondergaard J (2002), DACE: A matlab kriging toolbox ,version 2.0, Technical Report IMM-TR-2002-12, Technical University of Denmark, Copenhagen, 2002. http://www2.imm.dtu.dk/projects/dace/dace.pdf

Luo JQ, Liu F, McBean I (2015) Turbine blade row optimization through endwall contouring by an adjoint method. J Propuls Power 31:505–518

Martin JD, Simpson TW (2005), Use of kriging models to approximate deterministic computer models,AIAA Journal, 43(4): 853-863. https://doi.org/10.2514/1.8650

Myers RH, Montgomery DC (2002) Response surface methodology: process and product optimization using designed experiments, 2nd edn. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New York

Myung IJ, Mark AP (1997) Applying Occam's razor in modeling cognition: a Bayesian approach. Psychon Bull Rev 4(1):79–95

Namura N, Shimoyama K, Obayashi S (2017) Kriging surrogate model with coordinate transformation based on likelihood and gradient. J Glob Optim 68(4):827–849

Neath AA, Joseph EC (2012) The Bayesian information criterion: background, derivation, and applications. Wiley Interdisc Rev: Comput Stat 4(2):199–203

Park C, Haftka RT, Kim NH (2017) Remarks on multi-fidelity surrogates. Struct Multidiscip Optim 55(3):1–22

Rasmussen CE and Williams CK (2006), Gaussian processes for machine learning, MIT Press, London. http://www.gaussianprocess.org/gpml/

Shan SQ, Wang GG (2010) Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. Struct Multidiscip Optim 41(2):219–241

Suzen YB, Huang PG (2005) Numerical simulation of unsteady wake/blade interactions in low-pressure turbine flows using an intermittency transport equation. J Turbomach 127(3):431–444

Viana FAC, 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 Y, Schutte J, Meeker J, Palliyaguru U, Kim NH, Haftka RT (2017) Predicting B-basis allowable at untested points from experiments and simulations of plates with holes. In: 12th world congress on structural and multidisciplinary optimization, Braunschweig, Germany. URL: https://www.researchgate.net/publication/318909364