Performance study of Gradient Enhanced Kriging

Selvakumar Ulaganathan · Ivo Couckuyt · Tom Dhaene · Joris Degroote · Eric Laermans

Received: date / Accepted: date

Abstract The use of surrogate models for approximating computationally expensive simulations has been on the rise for the last two decades. Kriging-based surrogate models are popular for approximating deterministic computer models. In this work, the performance of Kriging is investigated when gradient information is introduced for the approximation of computationally expensive black-box simulations. This approach, known as Gradient Enhanced Kriging, is applied to various benchmark functions of varying dimensionality (2D-20D). As expected, results from the benchmark problems show that additional gradient information can significantly enhance the accuracy of Kriging. Gradient Enhanced Kriging provides a better approximation even when gradient information is only partially available. Further comparison between Gradient Enhanced Kriging and an alternative formulation of Gradient Enhanced Kriging, called indirect Gradient Enhanced Kriging, highlights various advantages of directly employing gradient information, such as improved surrogate model accuracy, better conditioning of the correlation matrix, etc. Finally, Gradient Enhanced Kriging is used to model 6- and 10-variable Fluid-Structure Interaction problems from bio-mechanics to identify the arterial wall's stiffness.

Keywords Kriging · Surrogate Modelling · Gradient Enhancement · Fluid Structure Interaction

S. Ulaganathan · I. Couckuyt · T. Dhaene · E. Laermans
Ghent University - iMINDS, Department of Information Technology (INTEC), Gaston Crommenlaan 8, 9050 Ghent, Belgium
E-mail: {selvakumar.ulaganathan, ivo.couckuyt, tom.dhaene, eric.laermans}@ugent.be

J. Degroote
Ghent University, Department of Flow, Heat and Combustion Mechanics, Sint - Pietersnieuwstraat 41, 9000 Ghent, Belgium
E-mail: Joris.Degroote@ugent.be

1 Introduction

The computational complexity of simulation codes, such as computational fluid dynamics (CFD) and finite element (FE) analysis, has grown rapidly in recent years, despite the continual advancement in computing power. In this respect, mimicking the behaviour of computationally intensive functions with simple approximations, known as surrogate models (surrogate modelling), has gained much attention among researchers over the past two decades. The aim of surrogate modelling is to accurately mimic the behaviour of a computationally expensive simulator over an input space of interest based on a limited number of expensive simulations (data points). To that end, it is crucial to take advantage of all additional available information, such as gradients, Hessian data, multi-fidelity data, prior knowledge, etc. [3,8,21,39,40]. For example, Kennedy and O'Hagan [14] investigated surrogate modelling techniques based on multi-fidelity data of variable computational cost. Yamazaki et al. [38] developed gradient and Hessian enhanced surrogate models with improved accuracy over models based on function data only. This is also the context of the work presented here; but, this work is more concerned with investigating the effect of a complete or partial set of gradient enhancement on the accuracy of the surrogate models.

Simpson et al. [30] and Wang et al. [37] provided an overview of various surrogate modelling techniques. Among them, Kriging, which was proposed by Sacks et al. [27] for the design and analysis of computer experiments, is very popular in computer aided engineering (CAE) applications to approximate deterministic data [16,26,29]. The modelling efficiency of Kriging is largely determined by the ability of its correlation function to capture the actual local behaviour of the function to be modelled. Hence, properties such as smoothness, differentiability, etc. of correlation functions play a significant role in the spatial interpolation abilities of Kriging. Morris et al. [23] proposed an extension of Kriging, called direct CoKriging method, which incorporates gradient information along with the function values to provide a more accurate approximation. As the term CoKriging is more used to describe multi-fidelity data modelling, we use the term direct Gradient Enhanced Kriging (GEK) instead to avoid confusion. Chung et al. [2] compared direct GEK with an alternative formulation of GEK, called indirect GEK, which uses the same mathematical formulation of Kriging, but augments the training data with additional function values estimated from the gradient information. The authors applied both GEK methodologies to an aerodynamic shape optimization problem and stated that both formulations are almost identical in performance. The authors further stated that indirect GEK is prone to numerical errors introduced during the estimation of additional function values from gradients whereas direct GEK exhibits formulation complexity at high dimensionality. Weiyu Liu [20] further investigated indirect GEK and proposed an alternative approach based on Neural Networks, trained with both function and gradient data, but its performance is lower than indirect GEK. Laurenceau and Sagaut [18] studied an aerodynamic problem with direct and indirect formulations of GEK and concluded that indirect GEK outperforms direct GEK irrespective of their equivalent mathematical formulations.

GEK has an extra requirement that the correlation function must be twice differentiable in order to calculate the correlation between gradient observations. An elaborate discussion on various differentiable correlation functions is given by various authors [24,25,31,36]. The Gaussian correlation structure is more commonly used in Kriging. However, Stein argues that the Gaussian correlation structure does not provide much flexibility in capturing the actual local behaviour of spatially varying quantities and encourages the use of the Matérn class of correlation functions due to its flexibility and manageable number of parameters [31].

In this paper, we are particularly interested in investigating the effect of gradient enhancement in Kriging with various correlation functions. To that end, the analytical expressions for the derivatives of various correlation functions with respect to design variables are derived, and the GEK methodology is also investigated if only a part of the gradient information is available. Based on the results of this investigation, a rule-of-thumb is proposed for the selection of GEK over Ordinary Kriging (OK) with respect to surrogate model accuracy and extra computational cost of estimating derivatives. In addition, a guideline is developed to improve the conditioning of the correlation matrix of GEK based on the tentative relationship between the values of the hyper-parameters and the accuracy of GEK models. Further, a guideline is suggested to reduce the surrogate model fitting cost of GEK by sorting the most relevant dimensions and only incorporating derivatives which correspond to those dimensions. Furthermore, the analytical equation of the likelihood gradients is derived and integrated in the GEK formulation, by evaluating the gradients of various correlation functions with respect to the hyper-parameters. This study is carried out by applying GEK to various benchmark functions of varying dimensionality (2D-20D) and to one real-life problem from bio-mechanics. This study captures and highlights the efficiency, applicability and limitations associated with Gradient Enhanced Kriging.

The remaining part of this paper is structured as follows: in Section 2, the mathematical formulation of Gradient Enhanced Kriging is presented. Section 3 discusses gradient incorporation in various correlation functions which are suitable for Gradient Enhanced Kriging. Section 4 lists the employed benchmark and the real-life problems. Test results are presented and discussed in Section 5. Finally, conclusions are drawn in Section 6.

2 Gradient Enhanced Kriging (GEK)

The mathematical form of a Kriging model has two parts as shown in Equation 1. The first part, $\hat{\mu}$, represents a trend function and the second part, which captures the local deviations from the trend function, is the realization of a stationary Gaussian random process.

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \boldsymbol{\psi}^T \boldsymbol{\Psi}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}), \tag{1}$$

where the column vector $\boldsymbol{\psi}$ contains the correlation between the sample data and a prediction point $\mathbf{x}^*$, $\mathbf{y}$ corresponds to the column vector of function values (a.k.a. *response* throughout this paper) from the sample data and $\boldsymbol{\Psi}$ is the correlation matrix which contains the correlation between the sampled data points. Various correlation functions can be employed based on available prior information about the underlying function to be modelled. This is discussed in Section 3.

In the case of GEK, the correlation matrix ($\boldsymbol{\Psi}$) in Equation 1 becomes a block matrix,

$$\dot{\boldsymbol{\Psi}} = \left\{ \begin{array}{cccccc} \boldsymbol{\Psi} & \frac{\partial \boldsymbol{\Psi}}{\partial x_1^{(i)}} & \cdots & \frac{\partial \boldsymbol{\Psi}}{\partial x_v^{(i)}} & \cdots & \frac{\partial \boldsymbol{\Psi}}{\partial x_k^{(i)}} \\ \frac{\partial \boldsymbol{\Psi}}{\partial x_1^{(j)}} & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_1^{(i)} \partial x_1^{(j)}} & \cdots & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_1^{(i)} \partial x_v^{(j)}} & \cdots & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_1^{(i)} \partial x_k^{(j)}} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial \boldsymbol{\Psi}}{\partial x_u^{(j)}} & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_1^{(j)} \partial x_u^{(i)}} & \cdots & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_u^{(i)} \partial x_v^{(j)}} & \cdots & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_u^{(i)} \partial x_k^{(j)}} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{\Psi}}{\partial x_k^{(j)}} & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_1^{(j)} \partial x_k^{(i)}} & \cdots & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_v^{(j)} \partial x_k^{(i)}} & \cdots & \frac{\partial^2 \boldsymbol{\Psi}}{\partial x_k^{(i)} \partial x_k^{(j)}} \end{array} \right\}, \tag{2}$$

where k is the dimensionality, i.e., number of design variables. Hence, Equation 1 for GEK becomes,

$$\hat{y}(\mathbf{x}^*) = \hat{\dot{\mu}} + \dot{\boldsymbol{\psi}}^T \dot{\boldsymbol{\Psi}}^{-1}(\dot{\mathbf{y}} - \mathbf{f}\hat{\dot{\mu}}), \tag{3}$$

where

$$\dot{\boldsymbol{\psi}} = \left(\boldsymbol{\psi}^T, \left(\frac{\partial \boldsymbol{\psi}}{\partial x_1} \right)^T, ..., \left(\frac{\partial \boldsymbol{\psi}}{\partial x_k} \right)^T \right)^T, \tag{4}$$

$$\dot{\mathbf{y}} = \left(\mathbf{y}^T, \left(\frac{\partial \mathbf{y}}{\partial x_1} \right)^T, ..., \left(\frac{\partial \mathbf{y}}{\partial x_k} \right)^T \right)^T, \tag{5}$$

$$\mathbf{f} = \left(1_1, ...1_{n_s}, 0_{n_s+1}, ..., 0_{(k+1)n_s} \right)^T, \tag{6}$$

where n_s is the number of sample points and the correlation vector $\dot{\boldsymbol{\psi}}$ contains the correlation of both function values and gradients between the sample data and the prediction point $\mathbf{x}^*$. The vector, $\dot{\mathbf{y}}$, contains both the function values and gradients of the sample data. The constant trend function for GEK, $\hat{\mu}$, is calculated via the general least square method as,

$$\hat{\mu} = (\mathbf{f}^T \dot{\mathbf{\Psi}}^{-1} \mathbf{f})^{-1} \mathbf{f}^T \dot{\mathbf{\Psi}}^{-1} \dot{\mathbf{y}}. \tag{7}$$

The hyper-parameters of the GEK model ($\theta_m, m = 1, ..., k$. See Section 3) can be obtained by maximizing the concentrated likelihood function,

$$\phi = \frac{-(k+1)n_s \; ln(\hat{\sigma}^2) - ln|\dot{\mathbf{\Psi}}|}{2}, \tag{8}$$

where $\hat{\sigma}^2$ is the estimated GEK variance which can be expressed as,

$$\hat{\sigma}^2 = \left(\frac{(\dot{\mathbf{y}} - \mathbf{f}\hat{\mu})^T \dot{\mathbf{\Psi}}^{-1} (\dot{\mathbf{y}} - \mathbf{f}\hat{\mu})}{(k+1)n_s} \right). \tag{9}$$

3 Correlation Functions

The choice of correlation functions is important in GEK, as the correlation functions must be differentiated once to calculate the correlation between response and gradient observations. Additionally, the correlation functions must be differentiated twice to calculate the correlation between gradient observations. Due to this fact, we limit ourselves to one stationary correlation function and two instances of the Matérn class of correlation functions.

A popular class of stationary correlation functions is defined as [8],

$$\psi(d) = exp\left(-\sum_{m=1}^{k} \theta_m d_m^{p_m} \right), \tag{10}$$

where $d = |x_m^i - x_m^j|$ and p is the smoothness parameter which defines the drop in correlation as d increases. A popular and more widely-used special case of the class of stationary correlation functions, which corresponds to setting $p = 2$, is the Gaussian correlation function [7,15,32]. The Gaussian correlation function assumes that the spatially varying function is smooth and continuous.

Stein [31] argues that such strong smoothness assumptions are unrealistic for modelling many physical processes, and recommends the Matérn class of correlation functions for its flexibility and highly recommendable spectral density. The flexibility is mainly due to its smoothness parameter (ν - similar to p in stationary correlation functions), as various values of ν guarantee good modelling accuracy for smooth surfaces (for $\nu \to \infty$, it becomes the Gaussian correlation function) as well as rough surfaces ($\nu \to 0$). Based on the value of ν, various instances of Matérn correlation functions are available. Matérn $\frac{3}{2}$ and Matérn $\frac{5}{2}$, which are more widely used in the machine learning context, can be described as,

$$\psi_{\nu=\frac{3}{2}}(d) = (1+\sqrt{3}a)exp\left(-\sqrt{3}a\right) \tag{11}$$

and

$$\psi_{\nu=\frac{5}{2}}(d) = (1+\sqrt{5}a+\frac{5a^2}{3})exp\left(-\sqrt{5}a\right), \tag{12}$$

where $a = \sqrt{\sum_{m=1}^{k} \theta_m d_m^2}$.

Instances of Matérn class of correlation functions are w times differentiable if and only if $\nu > w$ [25,31]. Hence, the Matérn $\frac{5}{2}$ instance is twice differentiable whereas the Matérn $\frac{3}{2}$ instance does not satisfy the differentiability requirement of GEK. But, recently Lockwood and Anitescu [21] used the Matérn $\frac{3}{2}$ correlation function in GEK and showed that the directional derivatives of the correlation function agree up to order 2, irrespective of the appearance of the absolute value in the correlation. Due to this nature and the highly rough behaviour of instances with $\nu < \frac{3}{2}$, the Matérn $\frac{3}{2}$ and $\frac{5}{2}$ correlation functions are used in this work to study the modelling performance of GEK for various test cases.

The gradient and the Hessian for the Gaussian correlation function are given by Equations 13 and 14, respectively.

$$\frac{\partial \mathbf{\Psi}^{(i,j)}}{\partial x_v^{(j)}} = 2\theta_v d_v \mathbf{\Psi}^{(i,j)} \tag{13}$$

$$\frac{\partial^2 \mathbf{\Psi}^{(i,j)}}{\partial x_u^{(i)} \partial x_v^{(j)}} = \begin{cases} -4\theta_u\theta_v d_u d_v \mathbf{\Psi}^{(i,j)} & \text{if } u \neq v \\ [2\theta - 4\theta^2 d^2]\mathbf{\Psi}^{(i,j)} & \text{if } u = v \end{cases} \tag{14}$$

The gradient and (somewhat convoluted form of) the Hessian for the Matérn $\frac{3}{2}$ correlation function are given by Equations 15 and 16, respectively.

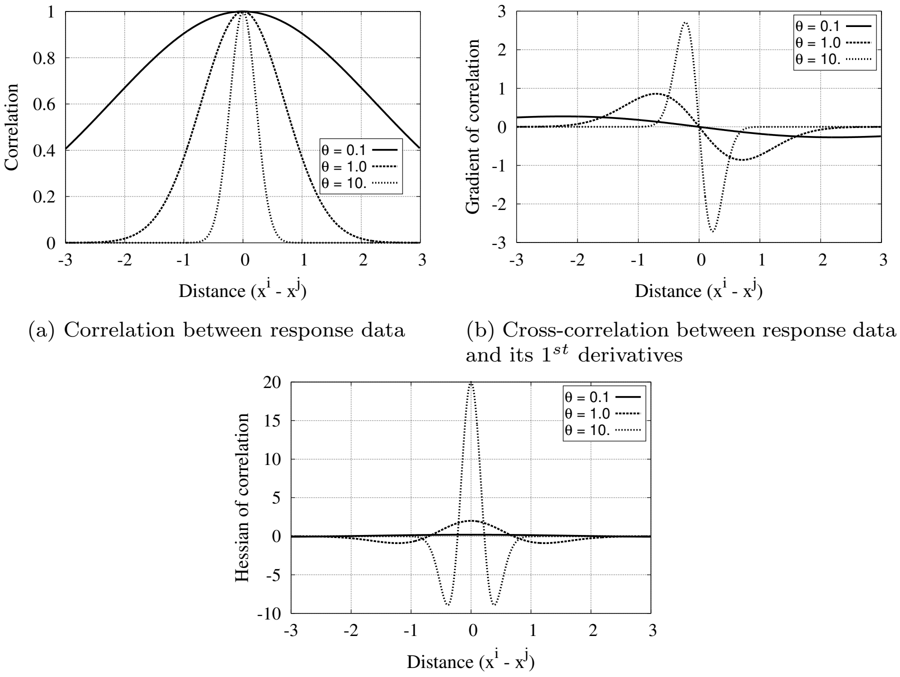
$$\frac{\partial \mathbf{\Psi}^{(i,j)}}{\partial x_v^{(i)}} = 3\theta_v d_v exp\left(-\sqrt{3}a\right) \tag{15}$$

$$\frac{\partial^2 \mathbf{\Psi}^{(i,j)}}{\partial x_u^{(i)} \partial x_v^{(j)}} = \begin{cases} \frac{-3\sqrt{3}\theta_u\theta_v d_u d_v exp\left(-\sqrt{3}a\right)}{a} & \text{if } u \neq v \\ 3\theta\left[1 - \frac{\sqrt{3}\theta d^2}{a}\right] exp\left(-\sqrt{3}a\right) & \text{if } u = v \end{cases} \tag{16}$$

The gradient and the Hessian for the Matérn $\frac{5}{2}$ correlation function are given by Equations 17 and 18, respectively.

$$\frac{\partial \mathbf{\Psi}^{(i,j)}}{\partial x_v^{(i)}} = \frac{5\theta_v d_v(\sqrt{5}a+1)exp\left(-\sqrt{5}a\right)}{3} \tag{17}$$

$$\frac{\partial^2 \mathbf{\Psi}^{(i,j)}}{\partial x_u^{(i)} \partial x_v^{(j)}} = \begin{cases} \frac{-25\theta_u\theta_v d_u d_v exp\left(-\sqrt{5}a\right)}{3} & \text{if } u \neq v \\ \left[\frac{-25\theta^2 d^2 + 5\theta(\sqrt{5}a+1)}{3}\right] exp\left(-\sqrt{5}a\right) & \text{if } u = v \end{cases} \tag{18}$$



(a) Correlation between response data

(b) Cross-correlation between response data and its 1^{st} derivatives

(c) Correlation between 1^{st} derivatives

Fig. 1: 1D correlation function (Gaussian)

Figures 1-3 show the influence of θ_k on the one dimensional correlation and cross-correlation functions of the Gaussian, Matérn $\frac{3}{2}$ and Matérn $\frac{5}{2}$ correlation functions. Figures 1a-3a show how far the influence of a sample point extends. Lower values of θ_k denote higher correlation among the sample points while the higher values denote that function values can change rapidly over a small region. Figures 1b-3b show how much deviation can happen from the surrogate model constructed with the first set of basis functions (i.e., correlation between the response values). Lower values of θ_k denote that larger areas are being influenced by the gradient value while the higher values denote smaller areas of distortion. The same kind of behaviour is depicted in Figures 1c-3c with twice differentiated correlation functions. For more elaborate information on how the choice of θ_k can influence the overall surrogate model accuracy, the reader is referred to [8].

The k dimensional non-linear optimization, which is performed to calculate suitable values of hyper-parameters, is the most time consuming part of Kriging. The optimization becomes even more computationally expensive in GEK as the correlation matrix has an additional $n_s \times k$ rows/columns. If the Cholesky decomposition is used to factorize $\dot{\mathbf{\Psi}}$, then the optimization requires a computational cost of $O(((k+1)n_s)^3)$ and a memory cost of $O(((k+1)n_s)^2)$. The Cholesky decomposition is the most expensive part of the optimization and, thus, the gradients of the likelihood with respect to the hyper-parameters θ_k are utilized to reduce the number of likelihood evaluations. Though the gradient estimation imposes an additional computational and memory cost, this may not be a significant issue for low-dimensional problems ($k \leq 20$). The likelihood gradients can either be calculated analytically or using reverse algorithmic differentiation of the likelihood. The latter is much faster and less dependent on the number of inputs [33]. In this work, the likelihood gradients for the GEK are calculated analytically by estimating the derivative of the concentrated likelihood function with respect to the hyper-parameters as [33],

$$\frac{\partial \phi}{\partial \theta} = \frac{1}{2\hat{\sigma}^2}\left[(\dot{\mathbf{y}} - \mathbf{f}\hat{\mu})^T \dot{\mathbf{\Psi}}^{-1}\frac{\partial \dot{\mathbf{\Psi}}}{\partial \theta}\dot{\mathbf{\Psi}}^{-1}(\dot{\mathbf{y}} - \mathbf{f}\hat{\mu})\right] - \frac{1}{2}\left[\dot{\mathbf{\Psi}}^{-1}\frac{\partial \dot{\mathbf{\Psi}}}{\partial \theta}\right]^T . \qquad (19)$$

Equation 19 can be solved once the derivatives of every element of $\dot{\mathbf{\Psi}}$ with respect to θ_k are calculated. As those derivatives are somewhat cumbersome to calculate, their analytical expressions are given in Appendix A. The derivative of the correlation and the cross-correlation functions with respect to θ_k for the Gaussian, Matérn $\frac{3}{2}$ and Matérn $\frac{5}{2}$ correlation functions are given by Equations 22 - 25, 26 - 30 and 31 - 36, respectively.

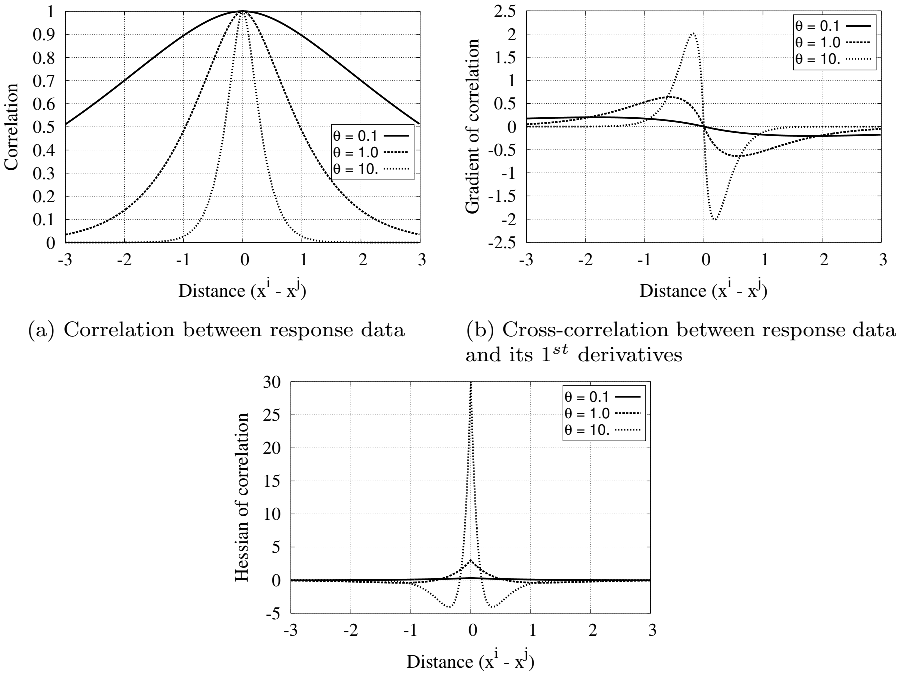
In this work, the k dimensional non-linear optimization is performed with Sequential Quadratic Programming (SQP) utilizing gradient information which is available in MATLAB[1] as the *fmincon* function. During the likelihood optimization, Equation 19 is solved for each hyper-parameter. As the analytical gradients of the likelihood function with respect to each hyper-parameter are passed to the optimizer along with the value of the likelihood function, the optimizer can quickly converge to the optimal values of the hyper-parameters with few evaluations of the likelihood function. This reduces the number of likelihood function evaluations, which in turn reduces the overall surrogate modelling time.

4 Problem Formulation

Four widely used benchmark functions[2] are employed as test functions, see Table 1. The gradient values of the benchmark functions with respect to the design variables are analytically calculated. A numerical simulator [5] that determines the difference between a given wall displacement and a calculated wall displacement for a given stiffness distribution along the length of an artery is used as a simulation example. Further details of this problem are given in Appendix B. The accuracy of resulting surrogate models is estimated with two different error measures: A validation data set and 'K-fold' Cross Validation (CV) [22], both using the Normalized Root Mean Square Error (NRMSE).

[1] MATLAB, The MathWorks, Inc., Natick, Massachusetts, USA

[2] www-optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar_files/TestGO_files/Page364.htm



(a) Correlation between response data

(b) Cross-correlation between response data and its 1^{st} derivatives

(c) Correlation between 1^{st} derivatives

Fig. 2: 1D correlation function (Matérn $\frac{3}{2}$)

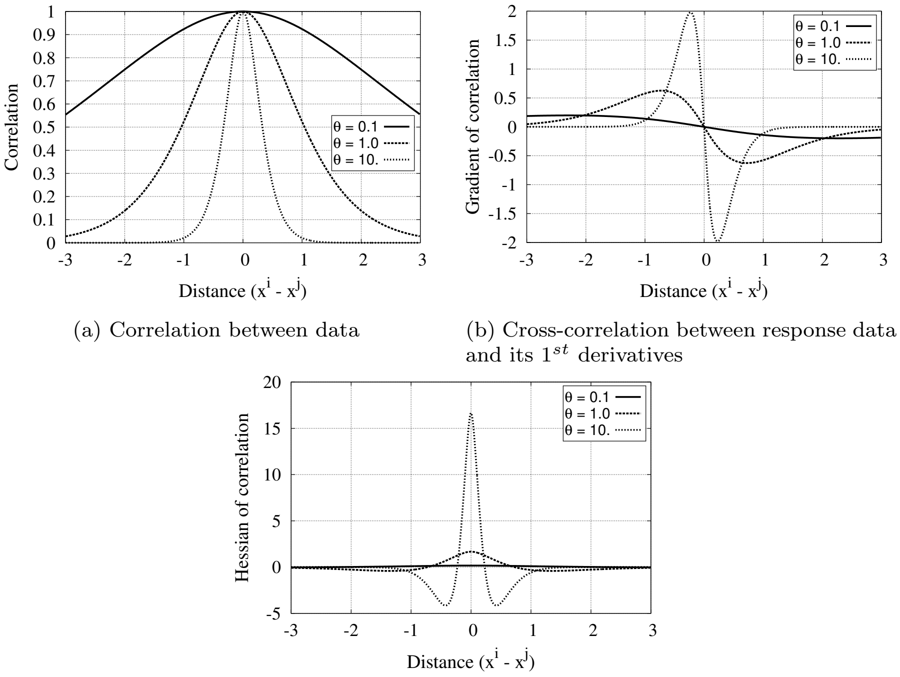
The validation data set contains 500 Monte Carlo points (n_p). The NRMSE for the response prediction can be defined as,

Table 1: Benchmark Test Problems

Benchmark Test Functions	Number of Dimensions	Properties
Fourhump camel back	2	Multi modal
Ackley	3	Multi modal
Hartmann	6	Multi modal
Rosenbrock	8	Unimodal and non-convex
Rosenbrock	20	Unimodal and non-convex

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n_p}\left(y_t^i - \hat{y}^i\right)^2}{n_p}}}{max(y_t) - min(y_t)}, \tag{20}$$

where y_t is the true response and $\hat{y}$ is the predicted response.



(a) Correlation between data

(b) Cross-correlation between response data and its 1^{st} derivatives

(c) Correlation between 1^{st} derivatives

Fig. 3: 1D correlation function (Matérn $\frac{5}{2}$)

Kriging surrogate models often require space-filling designs [16]. In this work, the class of Audze-Eglais Latin hypercube designs is considered to generate sample points [13]. The Audze-Eglais Latin hypercube designs provided by [13] supports up to a maximum of 300 sample points. Hence, the class of maximin Latin hypercube designs available in MATLAB is used to generate samples ranging from 100 to 1000. Moreover, the uniformly distributed pseudorandom designs available in MATLAB are also used to investigate the evolution of NRMSE, which is averaged over 50 independent runs for each sample size, with respect to the number of sample points (n_s).

5 Results and Analysis

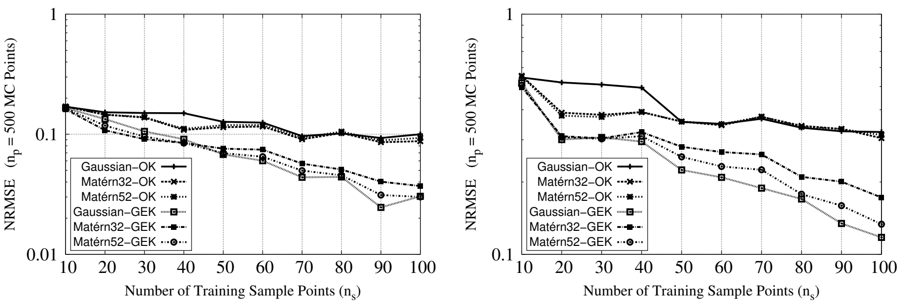
5.1 Benchmark Test Problems

Figures 4-7 show the evolution of NRMSE as a function of the number of training samples (n_s) for the benchmark test functions. As expected, the performance deviation of GEK over Ordinary Kriging (OK) is more pronounced with increasing sample size and dimensionality (k). This is essentially due to the fact that at a given n_s, GEK incorporates more information in the form of derivatives as the dimensionality of the problem increases. Although one can expect that both GEK and OK will converge to a similar accuracy level once a sufficient number of sample points is reached to model the underlying function, it is often difficult to know the appropriate size of the training data a priori. Though no general rule can be extracted about the required sample size for GEK to reach the accuracy level of OK as it is based on the complexity of the underlying function, a considerable sample reduction (50% - 90%) is achieved by GEK. The reduction in sample size achieved with GEK, in terms of % of reduction in sample size, to reach the same accuracy level of OK with $n_s = 100$ is given in Table 2.

In order to understand the robustness of GEK with various correlation functions, benchmark test functions are modelled using random designs instead of optimal Audze-Eglais Latin hypercube designs (Figures 8-10). For each number of training samples, ranging from 10 to 100 samples in steps of 10 samples, 50 random designs are constructed. Thus, the fitting of the OK and the GEK models is repeated 50 times for each training sample size. Almost an equal degree of improvement in performance over the OK models is achieved by the GEK models with different correlation functions. No single correlation function is observed to completely outperform its counterparts. However, the Gaussian and the Matérn $\frac{5}{2}$ correlation functions are the most consistent. Moreover, it is important to note that the Matérn $\frac{3}{2}$ correlation function doesn't deviate much from its near (Matérn $\frac{5}{2}$ correlation function) and extreme (the Gaussian correlation function when $\nu \to \infty$) counterparts irrespective of the conservative second-order differentiability nature as explained in Section 3. For more information on the suitability of various correlation functions in Kriging-based surrogate modelling, the reader is referred to [9–11]. Tables 3, 4, 14 (Appendix C) and 15 (Appendix C) give the CVE measure for the benchmark functions and the improvement in surrogate model accuracy reached by the GEK models against the Gaussian correlation function based OK models.

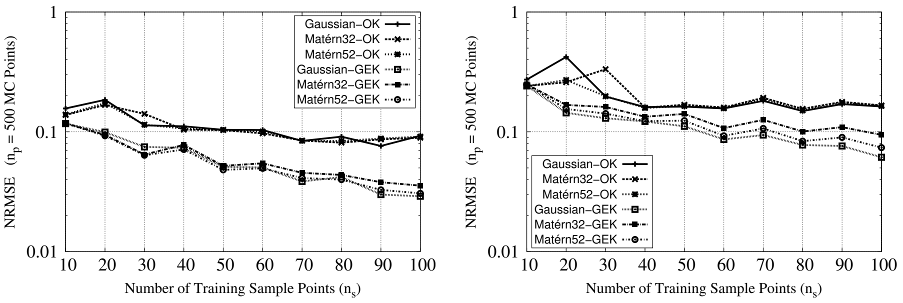
Table 2: Reduction in n_s with GEK (NRMSE on validation data set)

Benchmark Function	% of reduction in n_s by GEK to reach the same accuracy level of OK with $n_s = 100$	
	Response Prediction	Derivative Prediction
Ackley-3D	$> 50\% \pm 5.34\%$	$> 50\% \pm 5.23\%$
Hartmann-6D	$> 50\% \pm 5.57\%$	$> 50\% \pm 5.25\%$
Rosenbrock-8D	$> 70\% \pm 7.75\%$	$> 70\% \pm 7.93\%$
Rosenbrock-20D	$> 60\% \pm 6.88\%$	$> 85\% \pm 6.07\%$



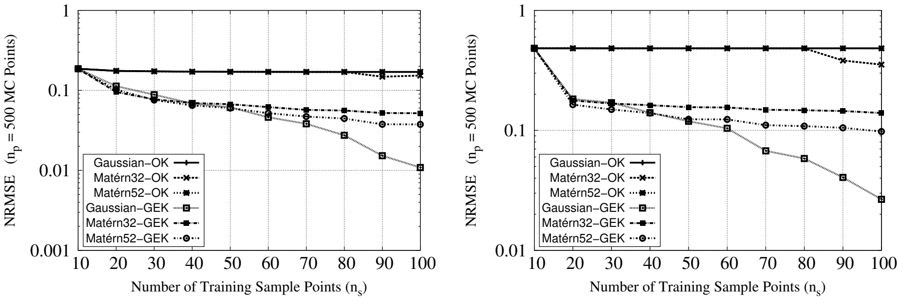
(a) Prediction of Response (b) Prediction of Derivatives

Fig. 4: Evolution of the NRMSE (Ackley-3D)



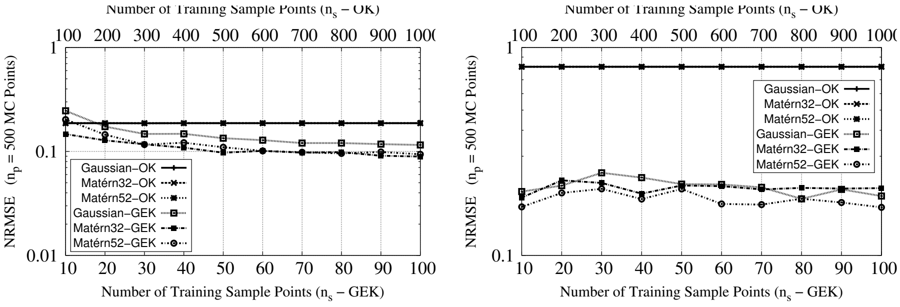
(a) Prediction of Response (b) Prediction of Derivatives

Fig. 5: Evolution of the NRMSE (Hartmann-6D)



(a) Prediction of Response (b) Prediction of Derivatives

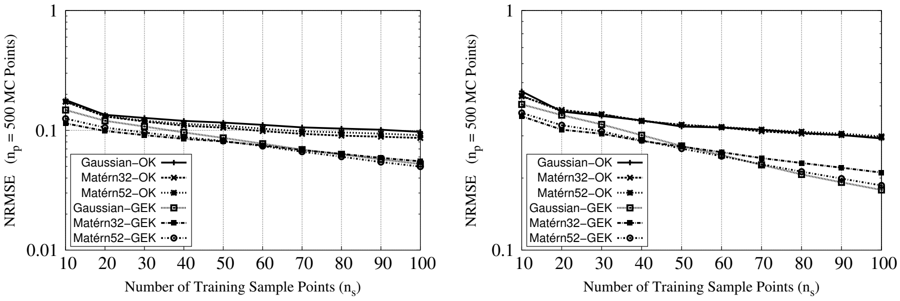
Fig. 6: Evolution of the NRMSE (Rosenbrock-8D)



(a) Prediction of Response

(b) Prediction of Derivatives

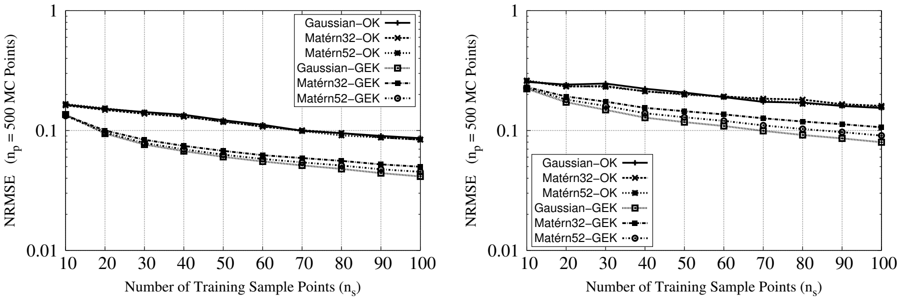
Fig. 7: Evolution of the NRMSE (Rosenbrock-20D)



(a) Prediction of Response

(b) Prediction of Derivatives

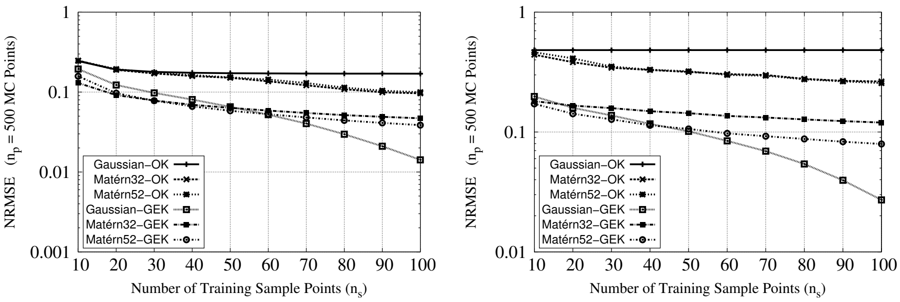
Fig. 8: Evolution of the NRMSE (averaged over 50 independent runs) (Ackley-3D)



(a) Prediction of Response

(b) Prediction of Derivatives

Fig. 9: Evolution of the NRMSE (averaged over 50 independent runs) (Hartmann-6D)



(a) Prediction of Response

(b) Prediction of Derivatives

Fig. 10: Evolution of the NRMSE (averaged over 50 independent runs) (Rosenbrock-8D)

Table 3: Prediction of Response

Benchmark Function		CV Error	[10 - fold]	[$n_s = 100/$	140(20D)]	
	Gaussian		Matérn $\frac{3}{2}$		Matérn $\frac{5}{2}$	
	OK	GEK	OK	GEK	OK	GEK
Ackley-3D	2.36e-01	1.24e-01	2.03e-01	1.28e-01	2.31e-01	1.16e-01
Hartmann-6D	3.02e-01	1.18e-01	2.82e-01	1.52e-01	2.89e-01	1.27e-01
Rosenbrock-8D	3.04e-01	2.62e-02	1.71e-01	8.47e-02	1.67e-01	7.09e-02
Rosenbrock-20D	3.23e-01	2.01e-01	3.23e-01	1.84e-01	3.23e-01	1.75e-01

Table 4: Efficiency of GEK and the Matérn class of correlation functions (Prediction of Response)

Benchmark Function		% of Improvement	[10 - fold]	[$n_s = 100/$	140(20D)]	
	Gaussian		Matérn $\frac{3}{2}$		Matérn $\frac{5}{2}$	
	OK	GEK	OK	GEK	OK	GEK
Ackley-3D	-	47% ± 2%	14% ± 2%	45% ± 2%	2% ± 2%	51% ± 2%
Hartmann-6D	-	61% ± 2%	7% ± 2%	50% ± 2%	4% ± 2%	58% ± 2%
Rosenbrock-8D	-	91% ± 2%	44% ± 2%	72% ± 2%	45% ± 2%	77% ± 2%
Rosenbrock-20D	-	38% ± 2%	0% ± 1%	43% ± 2%	0% ± 1%	46% ± 2%

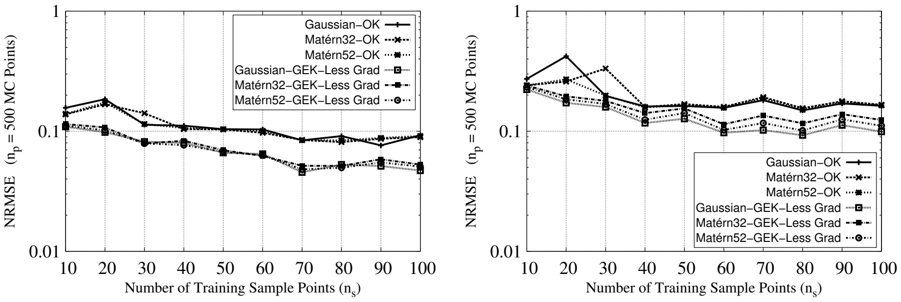
5.2 Effect of using partial set of Gradients

The accuracy of GEK models is assessed by leaving out gradient information in some of the dimensions of the training sample points during modelling. This can give an insight on how the GEK methodology will perform when the gradient information is only partially available. In addition, this also reduces the size of $\dot{\mathbf{\Psi}}$ to $(n_s + (n_s \times k')) \times (n_s + (n_s \times k'))$ with k' being the number of dimensions in which the partial set of gradients is incorporated. The GEK methodology is observed to result in more accurate surrogate models than OK even when the gradient information for some of the dimensions of the function to be modelled is completely left out (Figures 11-16). The dynamics of the left out gradients influence the surrogate model accuracy. For a given n_s, a minimum number of gradients in single dimension is required to produce GEK models which are more accurate than OK models for the Hartmann 6D function. Although it is based on the complexity of the function being modelled, this feature offers the possibility of improving the conditioning of $\dot{\mathbf{\Psi}}$ by leaving the least influencing gradients. Tables 5-8 give the improvement in surrogate model accuracy achieved by the GEK models, which are built with only the partial set of gradients in the specified dimension, over the Gaussian correlation function based OK models. Further, the cumulative accuracy improvement achieved by the GEK models when gradients in more than one dimension are successively incorporated is given by the last column in Tables 5-8. It should be noted that the sum of improvements of the GEK models of individual derivatives does not correspond to the improvement found in the GEK model incorporating all those derivatives. This can be explained by the different optimal hyper-parameter values found in the GEK model incorporating all those derivatives. The order of dimensions in Tables 5-8 corresponds to the order of values of the hyper-parameters. The values of the hyper-parameters in Tables 5-8 are obtained with the Gaussian correlation function based GEK models with $n_s = 100$. It can be observed that only incorporating gradients in the highest θ valued dimension shows significant improvement in the accuracy of GEK models. This means that accurate GEK models can be constructed by only incorporating partial set of gradients which corresponds to the largest value of θ [3]. In addition, the GEK model fitting cost can be considerably reduced by sorting the most relevant dimensions and only incorporating partial set of gradients which corresponds to those dimensions, as a GEK model with gradients in all the dimensions might become infeasible for high dimensional problems due to enormous size of $\dot{\mathbf{\Psi}}$. From the benchmark results, only including partial set of gradients in $\frac{k}{2}$ dimensions is observed to be a good trade-off between the GEK model fitting cost and model accuracy. Moreover, the accuracy of GEK models with $\frac{k}{2}$ partial set of gradients is not too far from that of GEK models with a complete k set of gradients (Tables 5-8).

5.3 Choosing GEK over OK

When the computational cost of acquiring derivative data becomes significant, it is important to select an appropriate technique among OK and GEK. For example, consider the Ackley 3D function for which the computational cost of estimating one function value and one set of k dimensional derivatives are t_y

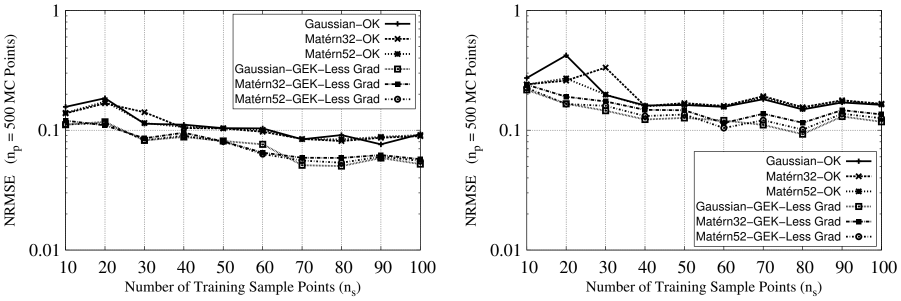
[3] This fact may not be completely true in high dimensional problems where partial set of gradients for a set of high-valued hyper-parameters is required to provide accurate GEK models (Table 8). Again, the size of the set of high-valued hyper-parameters, which is greater than 5 in this case, depends on the complexity of the function to be modelled.



(a) Prediction of Response

(b) Prediction of Derivatives

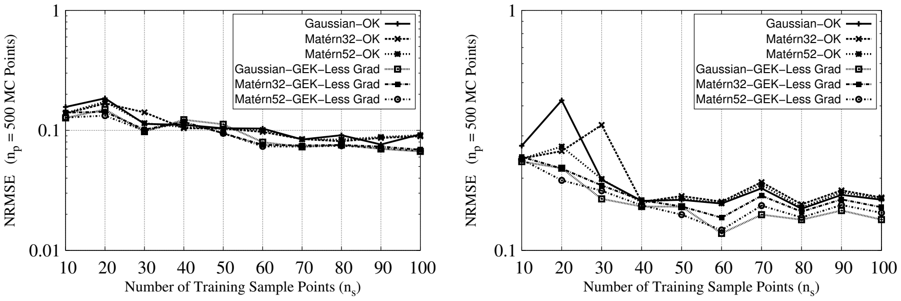
Fig. 11: Evolution of the NRMSE when 1^{st} dimension gradients are left out completely (Hartmann-6D)



(a) Prediction of Response

(b) Prediction of Derivatives

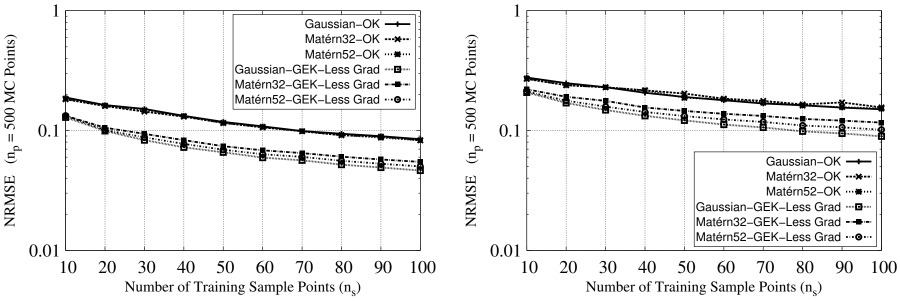
Fig. 12: Evolution of the NRMSE when 1^{st}-3^{rd} dimension gradients are left out completely (Hartmann-6D)



(a) Prediction of Response

(b) Prediction of Derivatives

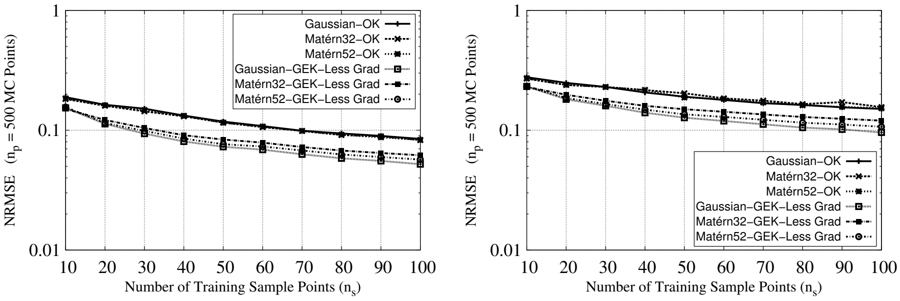
Fig. 13: Evolution of the NRMSE when 1^{st}-5^{th} dimension gradients are left out completely (Hartmann-6D)



(a) Prediction of Response

(b) Prediction of Derivatives

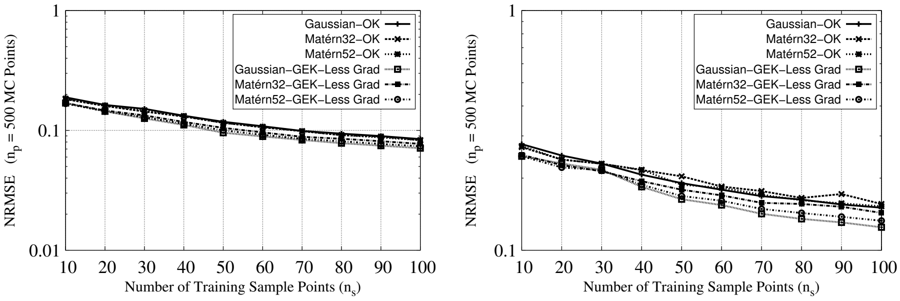
Fig. 14: Evolution of the NRMSE (averaged over 50 independent runs) when 1^{st} dimension gradients are left out completely (Hartmann-6D)



(a) Prediction of Response

(b) Prediction of Derivatives

Fig. 15: Evolution of the NRMSE (averaged over 50 independent runs) when 1^{st}-3^{rd} dimension gradients are left out completely (Hartmann-6D)



(a) Prediction of Response

(b) Prediction of Derivatives

Fig. 16: Evolution of the NRMSE (averaged over 50 independent runs) when 1^{st}-5^{th} dimension gradients are left out completely (Hartmann-6D)

Table 5: Efficiency of the Gaussian correlation function based GEK models with partial set of gradients (Ackley-3D: Prediction of Response - CVE with $n_s = 100$ and $n_{fold} = 10$)

Dimension	value of θ (Gaussian)	% of Improvement (over OK)	% of cumulative Improvement (over OK)
1	2.38	31%	31%
2	2.36	28%	40%
3	2.26	28%	47%

Table 6: Efficiency of the Gaussian correlation function based GEK models with partial set of gradients (Hartmann-6D: Prediction of Response - CVE with $n_s = 100$ and $n_{fold} = 10$)

Dimension	value of θ (Gaussian)	% of Improvement (over OK)	% of cumulative Improvement (over OK)
1	7.56	24%	24%
6	7.06	20%	27%
5	5.78	35%	42%
4	5.21	10%	49%
2	2.89	16%	60%
3	1.79	10%	61%

Table 7: Efficiency of the Gaussian correlation function based GEK models with partial set of gradients (Rosenbrock-8D: Prediction of Response - CVE with $n_s = 100$ and $n_{fold} = 10$)

Dimension	value of θ (Gaussian)	% of Improvement (over OK)	% of cumulative Improvement (over OK)
1	0.0351	51%	51%
4	0.0351	58%	60%
6	0.0348	55%	72%
5	0.0344	61%	76%
2	0.0344	57%	80%
3	0.0343	48%	81%
7	0.0341	46%	88%
8	0.0100	43%	91%

and t_{dy}, respectively. If $t_{dy=50} > 2 \times t_{y=50} (\approx t_{y=100})$, it is appropriate to choose OK over GEK, as it provides surrogate models with comparable accuracy to GEK without extra computational cost (Figure 4). On the contrary, if $t_{dy=50} < 2 \times t_{y=50} (\approx t_{y=100})$, then GEK would be an appropriate choice as it provides more accurate surrogate models than OK with negligible extra computational cost (Figure 4). Considering situations of this nature, a rule-of-thumb is proposed based on the results of the benchmark functions in order to tentatively guide one to select appropriately among OK and GEK.

Table 8: Efficiency of the Matérn $\frac{5}{2}$ correlation function based GEK models with partial set of gradients (Rosenbrock-20D: Prediction of Response - CVE with $n_s = 140$ and $n_{fold} = 10$)

Dimension	value of θ (Gaussian)	% of cumulative Improvement (over OK)
5-12-6-18-1	0.076-0.062-0.039-0.037-0.035	$< 1\%$
2-19-9-8-7	0.035-0.032-0.028-0.027-0.026	39%
4-13-14-10-11	0.026-0.025-0.024-0.024-0.023	39%
3-15-16-17-20	0.022-0.022-0.020-0.019-0.010	46%

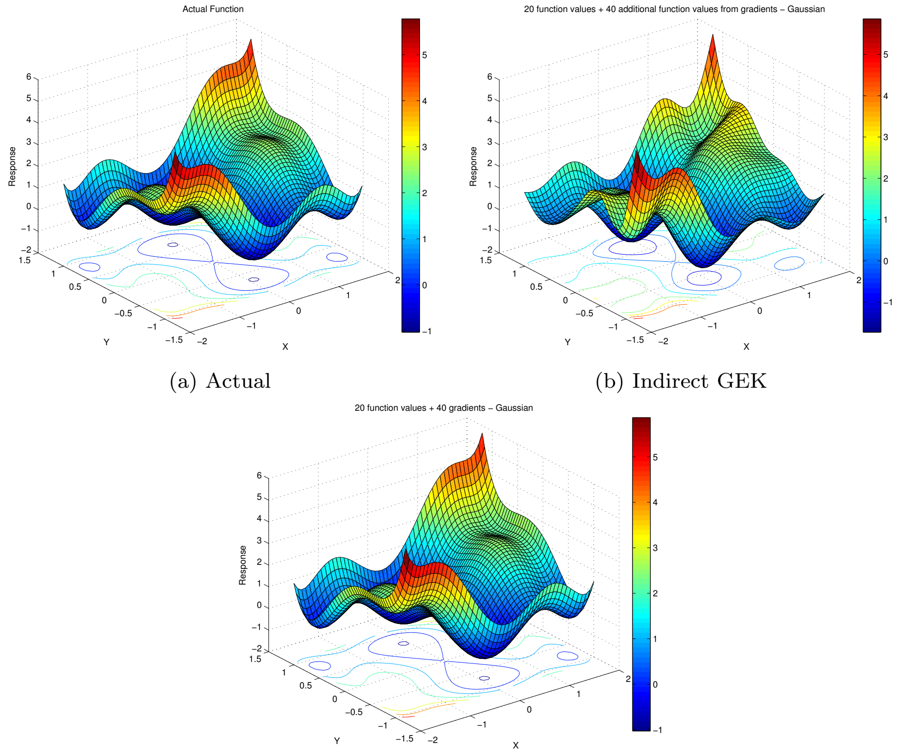
$$Model = \begin{cases} OK & \text{if } t_{dy} > \rho t_y \\ GEK & \text{if } t_{dy} < \rho t_y \qquad \rho \rightarrow 2 \quad to \quad 5, \end{cases} \tag{21}$$

where ρ is a constant relating to the dimensionality of the function to be modelled. In general, ρ is either 2 or 3 for functions with $k \leq 10$ and varies between 4 and 5 for functions with $10 < k \leq 20$. If the gradients are estimated by using an adjoint formulation based approach, the computational cost of evaluating the gradients is independent of the dimension of the problem and is directly proportional to the computational cost of a function evaluation. The value of ρ, in this case, depends on the details of the adjoint implementation and the worst-case estimate for ρ is 5, with typical values varying between 2 and 3 [12].

5.4 Direct and Indirect GEK

Gradient information can also be used in an indirect way, as shown by Chung et al. [2], where gradient information is used to estimate the nearby function values and correspondingly a standard OK model is built (Indirect GEK). Figure 17 depicts the 2D Fourhump camel back function which is approximated by the Direct GEK with 20 function and 40 gradient values whereas the Indirect GEK utilizes 20 function and 40 additional function values calculated from the available 40 gradient values. The additional function values are calculated based on a first-order Taylor approximation.

As expected, the indirect formulation struggles to capture the overall pattern of the underlying function while the direct formulation is able to model the quickly varying features more accurately as the gradient information is directly available. This is mainly caused by the complexities involved in estimating the right values for the interval in the first-order Taylor formula while calculating the additional function values for the indirect formulation. Moving too close or too far from the location of the original sample points results in no improvement or even a degradation in the accuracy of the indirect GEK models. As the direct formulation directly uses the gradient information, it tends



(a) Actual

(b) Indirect GEK

(c) Direct GEK

Fig. 17: Direct GEK vs Indirect GEK based on 20 samples and gradients (Fourhump-2D)

to be more accurate at higher n_s irrespective of the problems associated with dimensionality. Moreover, gradients estimated using finite-difference methods can also be incorporated in direct GEK. However, the computational cost of estimating gradients using finite-difference methods is very high as it takes k function evaluations to estimate gradient values at one sample point. Hence, the direct GEK with gradients estimated from finite-difference methods is not computationally advantageous than the indirect GEK with augmented function values. However, adjoint based-methods can be used to estimate gradients at very low computational cost [1,2,6,12,17,28].

5.5 Fluid Structure Interaction Problem

Figures 18-21 show the evolution of the NRMSE (and the averaged NRMSE over 50 independent runs) as a function of the number of training samples (n_s) for the FSI problems. A significant 80% - 90% reduction in sample size is achieved with GEK (Table 9). The feasibility of the sample reduction can be assessed by explaining the rule-of-thumb proposed in Section 5.3 and the extra computational cost associated with estimating derivatives. Table 10 gives the computational cost of estimating the function and derivative data for the 6D FSI function using a contemporary Ubuntu Linux desktop with $4 \times 2.10GHz$. As per Equation 21, t_{dy} should be greater than at least $2 \times t_y$ in order to choose OK over GEK. It can be observed from Table 10 that in no case t_{dy} is greater than $2 \times t_y$. Hence, it is more obvious to go for GEK over OK in this case. This choice can be further validated with the accuracy of the resulting GEK models. Table 10 shows that 97 function values can be calculated in the time of calculating 60 function and 6×60 derivative data for the 6D FSI function. When OK models of 97 function values are compared with the GEK models of 60 function and 6×60 derivative data, the GEK models outperform the OK models (Figures 18 and 20). A similar behaviour is exhibited for the 10D FSI function too. The extra computational cost associated with derivative estimation is quite cheap in the current case, thus making these comparisons even more tilted in favor of GEK.

However, for a fair comparison, the surrogate model fitting cost should be taken into account as well. As the derivative information is incorporated in GEK, the size of $\dot{\mathbf{\Psi}}$ grows substantially, and, hence, the surrogate model fitting cost. Hence, the training data of the OK models should be scaled to account for the additional derivative information incorporated in the GEK models. This is carried out by evaluating more function values and adding them to the correlation matrix of OK models so that the size of $\mathbf{\Psi}$ equals the size of $\dot{\mathbf{\Psi}}$. This leads to an equal surrogate model fitting cost for both OK and GEK models. Investigations on FSI problems show that OK with scaled training data results in more accurate surrogate models than GEK (Figure 22). This result goes along with the fact that a function value is more informative than a gradient value. However, it should be noted that the computational burden incurred by the additional function evaluations involved in OK models is significantly higher than the computational cost of acquiring function and derivative data for the GEK models with $\dot{\mathbf{\Psi}}$ being equal in size with $\mathbf{\Psi}$.

In addition, incorporating gradient information significantly improves the accuracy of GEK models, particularly in functions having $k \geq 8$, whereas the OK models show very small improvement in the accuracy as the number of response observation increases (Tables 11 and 16 (Appendix C)). Further, it can be seen in Tables 11 and 16 that a GEK model with $n_s = 20$ is more accurate than a OK model with $n_s = 100$. This may due to the fact that the hyper-parameter optimization of the OK models fails because of the small sample size in such a large design space. Hence the accuracy of all the OK models are almost the same. But, the gradients restrict the possible interpolation through the response data and allows the GEK models to successfully optimize its hyper-parameters [34]. This may actually be an advantage of GEK although in theory a response is worth more than a derivative. This can illustrate the feasibility of employing gradient values directly irrespective of the complexities involved in solving the large correlation matrix. Tables 12 and 17 (Appendix C) give the CVE measure for the FSI functions against varying dimensionality. Again, no single correlation function completely outperforms its counterparts.

Table 9: Reduction in n_s with GEK (NRMSE on validation data set)

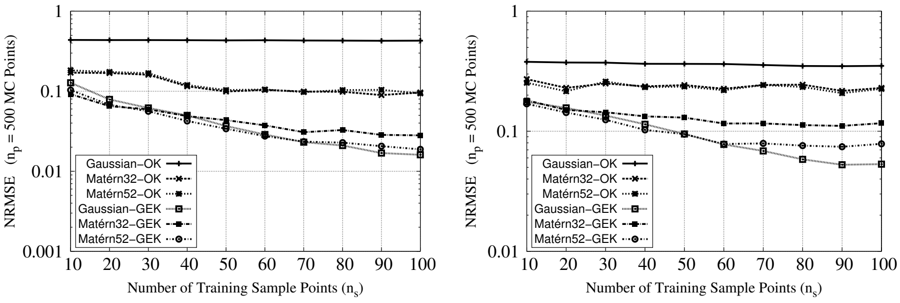
FSI Function	% of reduction in n_s by GEK to reach the same accuracy level of OK with $n_s = 100$	
	Response Prediction	Derivative Prediction
6D	$> 70\% \pm 8.9\%$	$> 80\% \pm 8.83\%$
10D	$> 50\% \pm 9.2\%$	$> 80\% \pm 9.9\%$

Table 10: Computational cost (6D FSI Function)

$n_s(GEK)$	t_y in s	t_{dy} in s	n_s (OK)
10	0.0010	0.0008	18
20	0.0018	0.0012	33
30	0.0027	0.0018	50
40	0.0037	0.0024	66
50	0.0042	0.0028	84
60	0.0055	0.0033	97

Table 11: Efficiency of GEK (10D FSI Function)

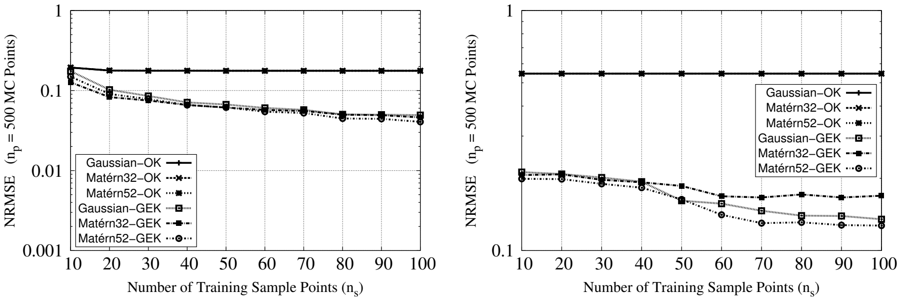
n_s		Response Prediction	(Matérn $\frac{5}{2}$)	
	OK		GEK	
	NRMSE	% of improvement	NRMSE	% of improvement
20	1.78e-01	–	9.1e-02	49% better
40	1.76e-01	0.8% better	6.6e-02	63% better
60	1.76e-01	0.9% better	5.4e-02	69% better
80	1.76e-01	1.0% better	4.5e-02	75% better
100	1.76e-01	1.0% better	4.1e-02	77% better



(a) Prediction of Response

(b) Prediction of Derivatives

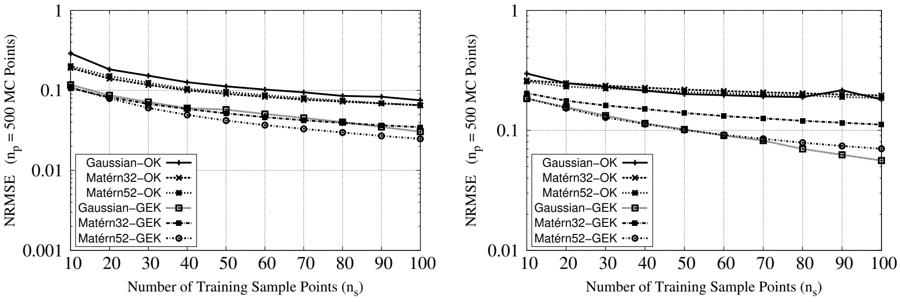
Fig. 18: Evolution of the NRMSE (6D FSI Function)



(a) Prediction of Response

(b) Prediction of Derivatives

Fig. 19: Evolution of the NRMSE (10D FSI Function)



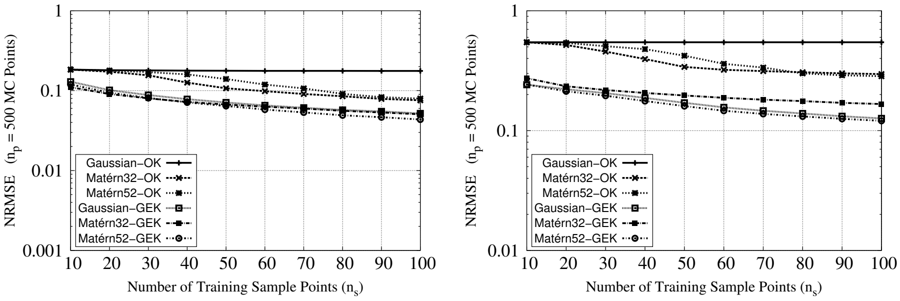
(a) Prediction of Response

(b) Prediction of Derivatives

Fig. 20: Evolution of the NRMSE (averaged over 50 independent runs) (6D FSI Function)

Table 12: Response Prediction

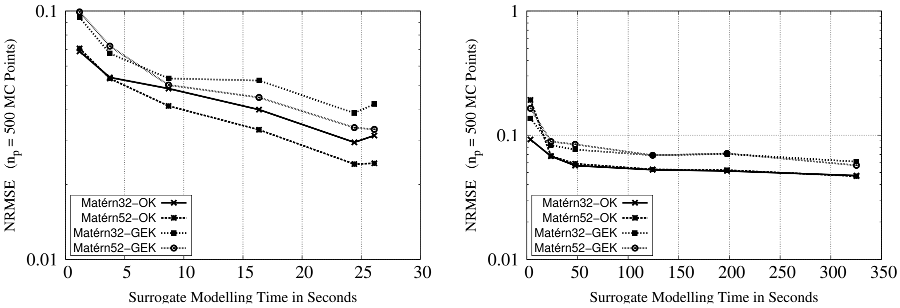
FSI Function		CV Error	[10 - fold]	$[n_s = 100]$		
	Gaussian		Matérn $\frac{3}{2}$		Matérn $\frac{5}{2}$	
	OK	GEK	OK	GEK	OK	GEK
6D	1.51e-01	5.66e-02	1.29e-01	6.71e-02	1.31e-01	5.10e-02
7D	2.67e-01	9.45e-02	1.81e-01	9.18e-02	1.84e-01	6.84e-02
8D	3.10e-01	9.96e-02	1.49e-01	8.25e-02	1.59e-01	7.74e-02
9D	3.08e-01	8.75e-02	1.51e-01	1.01e-01	1.59e-01	8.91e-02
10D	3.07e-01	9.17e-02	1.46e-01	7.94e-02	1.54e-01	7.52e-02



(a) Prediction of Response

(b) Prediction of Derivatives

Fig. 21: Evolution of the NRMSE (averaged over 50 independent runs) (10D FSI Function)



(a) Prediction of Response (FSI-6D)

(b) Prediction of Response (FSI-10D)

Fig. 22: Evolution of the NRMSE when the surrogate model fitting cost is equal for both OK and GEK. The size of the OK ‘correlation matrix’ is augmented with additional function values to equal that of the GEK ‘correlation’ matrix.

6 Conclusions

This paper investigates the effects of gradient enhancement in Kriging-based surrogate modelling. As expected, the gradient enhancement significantly reduces the number of training samples required to provide more accurate model representations. Based on the results of the investigation, a rule-of-thumb is proposed to make an appropriate choice among Ordinary Kriging and Gradient Enhanced Kriging when the computational cost of acquiring derivative data becomes dominant. Further, a tentative relationship between the values of the hyper-parameters and the accuracy improvement of Gradient Enhanced Kriging models with partial set of gradients is observed. Based on this tentative relationship, a guideline is developed to improve the conditioning of the correlation matrix, as this feature enables Gradient Enhanced Kriging to reduce the size of the correlation matrix and, subsequently the surrogate model fitting cost by discarding the least important gradients. In addition, a guideline is proposed for a good trade-off between the Gradient Enhanced Kriging model fitting cost and model accuracy. Furthermore, when the size of the correlation matrix of Ordinary Kriging is scaled with more function values in order to equal that of Gradient Enhanced Kriging (i.e., equal surrogate model fitting cost), Ordinary Kriging outperforms Gradient Enhanced Kriging; but, at a computational cost of estimating additional function values which is often higher than that of estimating derivatives for Gradient Enhanced Kriging. Although the direct and indirect formulations of Gradient Enhanced Kriging appear to be similar in performance, the direct formulation exhibits better conditioning of the correlation matrix. The use of analytical expressions for the likelihood gradients speeds up the overall hyper-parameters calculation time by reducing the number of function evaluations.

Acknowledgements This research has been funded by the Interuniversity Attraction Poles Programme BESTCOM initiated by the Belgian Science Policy Office. Additionally, this research has been supported by the Fund for Scientific Research in Flanders (FWO-Vlaanderen). Ivo Couckuyt and Joris Degroote are post-doctoral research fellows of the Research Foundation Flanders (FWO).

References

1. Brezillon, J., Dwight, R.: Discrete adjoint of the Navier-Stokes equations for aerodynamic shape optimization. In: Evolutionary and Deterministic Methods for Design, Optimisation and Control with Applications to Industrial and Societal Problems (EUROGEN 2005). Munich, Germany (2005)
2. Chung, H.S., Alonso, J.J.: Using gradients to construct cokriging approximation models for high-dimensional design optimization problems. In: Problems, 40th AIAA Aerospace Sciences Meeting and Exhibit, AIAA, pp. 2002–0317. Reno, NV (2002)
3. Couckuyt, I., Dhaene, T., Demeester, P.: ooDACE toolbox: A flexible object-oriented kriging implementation. Journal of Machine Learning Research **15**, 3183–3186 (2014)
4. Degroote, J., Bathe, K.J., Vierendeels, J.: Performance of a new partitioned procedure versus a monolithic procedure in fluid-structure interaction. Computers & Structures **87**(11–12), 793–801 (2009)
5. Degroote, J., Hojjat, M., Stavropoulou, E., Wüchner, R., Bletzinger, K.U.: Partitioned solution of an unsteady adjoint for strongly coupled fluid-structure interactions and application to parameter identification of a one-dimensional problem. Structural and Multidisciplinary Optimization **47**(1), 77–94 (2013)
6. Dwight, R.P., Han, Z.H.: Efficient uncertainty quantification using gradient-enhanced kriging. In: 11th AIAA Non-Deterministic Approaches Conference. Palm Springs, California, USA (2009)
7. Forrester, A.I., Sóbester, A., Keane, A.J.: Multi-fidelity optimization via surrogate modelling. Proceedings of the Royal Society **463**, 3251–3269 (2007)
8. Forrester, A.I., Sóbester, A., Keane, A.J.: Engineering Design via Surrogate Modelling: A Practical Guide, 1 edn. Wiley (2008)
9. Ginsbourger, D.: Multiples métamodéles pour l'approximation et l'optimisation de fonctions numériques multivariables. Ph.D. thesis, École des Mines de Saint-Étienne (2009)
10. Ginsbourger, D., Dupuis, D., Badea, A., Carraro, L., Roustant, O.: A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments. Appl. Stoch. Models Bus. Ind. **25**(2), 115–131 (2009). Special Issue: Computer Experiments versus Physical Experiments
11. Ginsbourger, D., Helbert, C., Carraro, L.: Discrete mixtures of kernels for kriging-based optimization. Quality and Reliability Eng. Int. **24**(6), 681–691 (2008)
12. Griewank, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. No. 19 in Frontiers in Appl. Math. SIAM, Philadelphia, PA (2000)
13. Husslage, B., Rennen, G., van Dam, E., den Hertog, D.: Space-filling latin hypercube designs for computer experiments. Optimization and Engineering **12**(4), 611–630 (2011)
14. Kennedy, M.C., O'Hagan, A.: Predicting the output from a complex computer code when fast approximations are available. Biometrika **87**(1), 1–13 (2000)
15. Killeya, M.R.H.: Thinking inside the box: using derivatives to improve Bayesian black box emulation of computer simulators with application to compart mental models. Ph.D. thesis, Durham theses, Durham University, England (2004)
16. Kleijnen, J.: Kriging metamodeling in simulation: A review. European Journal of Operational Research **192**(3), 707–716 (2009)
17. Laurenceau, J., Meaux, M., Montagnac, M., Sagaut, P.: Comparison of gradient-based and gradient-enhanced response-surface-based optimizers. American Institute of Aeronautics and Astronautics Journal **48**(5), 981–994 (2010)
18. Laurenceau, J., Sagaut, P.: Building efficient response surfaces of aerodynamic functions with kriging and cokriging. AIAA **46**(2), 498–507 (2008)

19. Laurent, S., Cockcroft, J., Van Bortel, L., Boutouyrie, P., Giannattasio, C., Hayoz, D., Pannier, B., Vlachopoulos, C., Wilkinson, I., Struijker-Boudier, H.: Expert consensus document on arterial stiffness: methodological issues and clinical applications. European Heart Journal **27**(21), 2588–2605 (2006)
20. Liu, W.: Development of gradient-enhanced kriging approximations for multidisciplinary design optimisation. Ph.D. thesis, University of Notre Dame, Notre Dame, Indiana (2003)
21. Lockwood, B.A., Anitescu, M.: Gradient-enhanced universal kriging for uncertainty propagation. Preprint ANL/MCS-P1808-1110 (2010)
22. Meckesheimer, M., Booker, A.J., Barton, R.R., Simpson, T.W.: Computationally inexpensive metamodel assessment strategies. AIAA **40**(10), 2053–2060 (2002)
23. Morris, M.D., Mitchell, T.J., Ylvisaker, D.: Bayesian design and analysis of computer experiments: Use of gradients in surface prediction. Technometrics **35**(3), 243–255 (1993)
24. Näther, W., Šimák, J.: Effective observation of random processes using derivatives. Metrika Springer-Verlag **58**, 71–84 (2003)
25. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. The MIT Press, Cambridge, MA, USA (2006)
26. Sacks, J., Schiller, S.B., Welch, W.J.: Designs for computer experiments. Technometrics **31**(1), 41–47 (1989)
27. Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Statistical Science **4**(4), 409–423 (1989)
28. Schneider, R.: Feins: Finite element solver for shape optimization with adjoint equations. In: Progress in industrial mathematics at ECMI 2010 Conference, pp. 573–580 (2012)
29. Shao, W., Deng, H., Ma, Y., Wei, Z.: Extended Gaussian kriging for computer experiments in engineering design. Engineering with Computers **28**(2), 161–178 (2012)
30. Simpson, T., Poplinski, J., Koch, P.N., Allen, J.: Metamodels for computer-based engineering design: Survey and recommendations. Engineering with Computers **17**(2), 129–150 (2001)
31. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York (1999)
32. Stephenson, G.: Using derivative information in the statistical analysis of computer models. Ph.D. thesis, University of Southampton, Southampton, UK (2010)
33. Toal, D.J., Forrester, A.I., Bressloff, N.W., Keane, A.J., Holden, C.: An adjoint for likelihood maximization. Proc R Soc A 8 **465**(2111), 3267–3287 (2009)
34. Ulaganathan, S., Couckuyt, I., Ferranti, F., Laermans, E., Dhaene, T.: Performance study of multi-fidelity gradient enhanced kriging, accepted paper. Structural and Multidisciplinary Optimization (2014)
35. Vignon-Clementel, I., Figueroa, C., Jansen, K., Taylor, C.: Outflow boundary conditions for 3D simulations of non-periodic blood flow and pressure fields in deformable arteries. Computer Methods in Biomechanics and Biomedical Engineering **13**(5), 625–640 (2010)
36. Šimák, J.: On experimental designs for derivative random fields. Ph.D. thesis, TU Bergakademie Freiberg, Freiberg, Germany (2002)
37. Wang, G.G., Shan, S.: Review of metamodeling techniques in support of engineering design optimization. Journal of Mechanical Design **129**(4), 370–380 (2006)
38. Yamazaki, W., Rumpfkeil, M.P., Mavriplis, D.J.: Design optimization utilizing Gradient/Hessian enhanced surrogate model. In: 28th AIAA Applied Aerodynamics Conference, AIAA paper 2010-4363. Chicago, Illinois, USA (2010)
39. Ying, X., JunHua, X., WeiHua, Z., YuLin, Z.: Gradient-based kriging approximate model and its application research to optimization design. SCIENCE CHINA Technological Sciences **52**(4), 1117–1124 (2009)
40. Zhao, D., Xue, D.: A multi-surrogate approximation method for metamodeling. Engineering with Computers **27**(2), 139–153 (2011)

A Analytical expressions for likelihood gradients

A.1 Gaussian correlation function:

Derivative of correlation function with respect to θ_k:

$$\frac{\partial}{\partial \theta_k}\left(\psi(d_{u(v)})\right) = -10^{\theta_{u(v)}} d_{u(v)}^2 log(10) exp\left(-\sum_{m=1}^{k} \theta_m d_m^2\right) \tag{22}$$

Derivatives of cross-correlation functions with respect to θ_k:

$$\frac{\partial}{\partial \theta_k}\left(\frac{\partial \mathbf{\Psi}^{(i,j)}}{\partial x_v^{(i)}}\right) = \begin{cases} 2d_v 10^{\theta_v} log(10) \mathbf{\Psi}^{(i,j)} \left[1 - 10^{\theta_k} d_k^2\right] & \text{if } v = k \\ 2d_v 10^{\theta_v} log(10) \mathbf{\Psi}^{(i,j)} \left[-10^{\theta_k} d_k^2\right] & \text{if } v \neq k \end{cases} \tag{23}$$

$$\frac{\partial}{\partial \theta_k}\left(\frac{\partial^2 \mathbf{\Psi}^{(i,j)}}{\partial x_u^{(i)} \partial x_v^{(j)}}\right) = \begin{cases} -4d_u d_v 10^{\theta_u} 10^{\theta_v} log(10) \mathbf{\Psi}^{(i,j)} \left[1 - 10^{\theta_k} d_k^2\right] & \text{if } u|v = k \\ 4d_u d_v d_k^2 10^{\theta_u} 10^{\theta_v} 10^{\theta_k} log(10) \mathbf{\Psi}^{(i,j)} & otherwise \end{cases} \tag{24}$$

$$\frac{\partial}{\partial \theta_k}\left(\frac{\partial^2 \mathbf{\Psi}^{(i,j)}}{\partial x_{u=v}^{(i)} \partial x_{u=v}^{(j)}}\right) = \begin{cases} log(10) \mathbf{\Psi}^{(i,j)} \left[2(10^{\theta}) + 4(10^{3\theta}) d^4 - 10(10^{2\theta}) d^2\right] & \text{if } (u = v) = k \\ -log(10) \mathbf{\Psi}^{(i,j)} 10^{\theta_k} d_k^2 \left[2(10^{\theta}) - 4(10^{2\theta}) d^2\right] & \text{if } (u = v) \neq k \end{cases} \tag{25}$$

A.2 Matérn $\frac{3}{2}$ correlation function:

Derivative of correlation function with respect to θ_k:

$$\frac{\partial}{\partial \theta_k}\left(\psi_{\nu=3/2}(d_{u(v)})\right) = -1.5(10^{\theta_{u(v)}}) log(10) d_{u(v)}^2 exp\left(-\sqrt{3}a\right) \tag{26}$$

Derivatives of cross-correlation functions with respect to θ_k:

$$\frac{\partial}{\partial \theta_k}\left(\frac{\partial \mathbf{\Psi}^{(i,j)}}{\partial x_v^{(i)}}\right) = \begin{cases} 3(10^{\theta_v}) d_v log(10) exp\left(-\sqrt{3}a\right) \left[1 - \frac{\sqrt{3} 10^{\theta_k} d_k^2}{2a}\right] & \text{if } v = k \\ 3(10^{\theta_v}) d_v log(10) exp\left(-\sqrt{3}a\right) \left[\frac{-\sqrt{3} 10^{\theta_k} d_k^2}{2a}\right] & \text{if } v \neq k \end{cases} \tag{27}$$

$$\frac{\partial}{\partial \theta_k}\left(\frac{\partial^2 \mathbf{\Psi}^{(i,j)}}{\partial x_u^{(i)} \partial x_v^{(j)}}\right) = \begin{cases} V_1 \left[\dfrac{-10^{\theta_k} d_k^2}{2a^3} + \dfrac{1}{a} - \dfrac{\sqrt{3} 10^{\theta_k} d_k^2}{2a^2}\right] & \text{if } u|v = k \\ V_1 \left[\dfrac{-10^{\theta_k} d_k^2}{2a^3} - \dfrac{\sqrt{3} 10^{\theta_k} d_k^2}{2a^2}\right] & otherwise \end{cases} \tag{28}$$

$$\frac{\partial}{\partial \theta_k}\left(\frac{\partial^2 \mathbf{\Psi}^{(i,j)}}{\partial x_{u=v}^{(i)} \partial x_{u=v}^{(j)}}\right) = \begin{cases} \dfrac{-V_1 10^{\theta_k} d_k^2 (1 + \sqrt{3}a)}{2a^3} + \dfrac{2V_1}{a} + V_2 \left[1 - \dfrac{\sqrt{3} 10^{\theta_k} d_k^2}{2a}\right] & \text{if } (u = v) = k \\ \dfrac{-V_1 10^{\theta_k} d_k^2 (1 + \sqrt{3}a)}{2a^3} - \dfrac{V_2 \sqrt{3} 10^{\theta_k} d_k^2}{2a} & \text{if } (u = v) \neq k \end{cases} \tag{29}$$

where

$$V_1 = -3\sqrt{3} 10^{\theta_u} 10^{\theta_v} d_u d_v log(10) exp\left(-\sqrt{3}a\right) \quad \& \quad V_2 = 3(10^{\theta_{u=v}}) log(10) exp\left(-\sqrt{3}a\right) \tag{30}$$

A.3 Matérn $\frac{5}{2}$ correlation function:

Derivative of correlation function with respect to θ_k:

$$\frac{\partial}{\partial\theta_k}\left(\psi_{\nu=5/2}(d_{u(v)})\right) = \frac{-\left(5+5\sqrt{5}a\right)10^{\theta}log(10)d^2_{u(v)}exp\left(-\sqrt{5}a\right)}{6} \tag{31}$$

Derivatives of cross-correlation functions with respect to θ_k:

$$\frac{\partial}{\partial\theta_k}\left(\frac{\partial\mathbf{\Psi}^{(i,j)}}{\partial x_v^{(i)}}\right) = \begin{cases} 10^{\theta_v}d_vC_2\left[C_1+\left(\frac{-25}{6}\right)10^{\theta_k}d_k^2\right] & \text{if } v=k \\ 10^{\theta_v}10^{\theta_k}d_vd_k^2\left(\frac{-25C_2}{6}\right) & \text{if } v\neq k \end{cases} \tag{32}$$

$$\frac{\partial}{\partial\theta_k}\left(\frac{\partial^2\mathbf{\Psi}^{(i,j)}}{\partial x_u^{(i)}\partial x_v^{(j)}}\right) = \begin{cases} \dfrac{-25C_2\left(1-\frac{\sqrt{5}10^{\theta_k}d_k^2}{2a}\right)10^{\theta_u}10^{\theta_v}d_ud_v}{3} & \text{if } u|v=k \\ \dfrac{C_225\sqrt{5}10^{\theta_u}10^{\theta_v}10^{\theta_k}d_ud_vd_k^2}{6a} & otherwise \end{cases} \tag{33}$$

$$\frac{\partial}{\partial\theta_k}\left(\frac{\partial^2\mathbf{\Psi}^{(i,j)}}{\partial x_{u=v}^{(i)}\partial x_{u=v}^{(j)}}\right) = \begin{cases} V_3+V_4 & \text{if } (u=v)=k \\ V_3 & \text{if } (u=v)\neq k \end{cases} \tag{34}$$

where

$$V_3 = \left[\left(\frac{25\sqrt{5}}{6a}\right)(10^{\theta u=v})^2(d_{u=v})^2 - \left(\frac{25}{6}\right)10^{\theta u=v}\right]C_2 10^{\theta k}d_k^2 \tag{35}$$

$$V_4 = \left(\frac{-50C_2(10^{\theta})^2d^2}{3}\right) + C_1C_2 10^{\theta}; \qquad C_1 = \left(\frac{5\sqrt{5}}{3}a + \frac{5}{3}\right); \qquad C_2 = log(10)exp\left(-\sqrt{5}a\right) \tag{36}$$

B Problem description (Fluid structure interaction problem)

In 2008, the World Health Organization reported that cardiovascular diseases are the leading cause of death in the world. Aortic stiffening has been linked to many (patho)physiological mechanisms and conditions. In all of them, the stiffness and mechanical properties of the aortic wall are altered. Consequently, non-invasive measurements of the arterial stiffness are needed and assessing the arterial stiffness should be part of the routine clinical diagnosis and follow-up procedures [19]. In this example, the stiffness distribution along the length of an artery is identified using a simplified numerical model. Previously, unconstrained quasi-Newton optimization with line search and a discrete adjoint solver for the calculation of the gradient was applied for this purpose [5].

The numerical model is one-dimensional in an axisymmetric (r, ϕ, z) coordinate system, as depicted in Figure 23. It consists of $k-1$ elastic segments, each with its own stiffness. Inside the artery, there is an incompressible blood flow. Furthermore, the interaction between this blood flow and the elastic wall is taken into account.

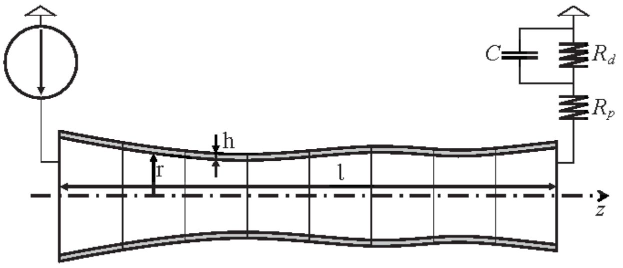


Fig. 23: The one-dimensional and axisymmetric model for blood flow in an artery with the prescribed velocity at the inlet (left) and the Windkessel model at the outlet (right). The segments, radius r, wall thickness h and length ℓ are indicated.

The blood flow rate at a point can be measured as a function of time using non-invasive techniques. So, the flow rate at the inlet is prescribed as a function of the time t with a period corresponding to one heart beat t_b. A Windkessel model relates this velocity with the outlet pressure [35]. This Windkessel model (See Figure 23) represents the remainder of the circulation, downstream from the artery. The capacitor C represents the compliance of the arterial system, while the resistors R_p and R_d model the proximal and distal viscous resistance, respectively.

The goal is to adjust the stiffness parameters of this fluid-structure interaction model so that the displacement of the arterial wall as a function of time matches the displacement data from a non-invasive measurement. The elasticity modulus E_i of each segment i ($i \in \{1, \ldots, k-1\}$) is modified by the corresponding parameter x_i which varies from -1 to 1.

$$E_i = E_o \left(1 + \frac{1}{2} x_i\right) \tag{37}$$

As the Windkessel model also has a significant impact on the wall displacement, the value of C is modified by the parameter x_k which varies from -1 to 1.

$$C = C_o \left(1 + \frac{1}{2} x_k\right)^{-1} \tag{38}$$

The parameter x_k will be identified, together with the parameters x_i ($i \in \{1, \ldots, k-1\}$). All fixed parameters are listed in Table 13.

Table 13: The parameters of the fluid-structure interaction model and the Windkessel model [35].

r_o	$3{\cdot}10^{-3}$ m	E_o	$4{\cdot}10^{5}$ Pa
h	$3{\cdot}10^{-4}$ m	C_o	$6.35{\cdot}10^{-10}$ m^3/Pa
ℓ	0.126 m	R_d	$1.768{\cdot}10^{9}$ Pa·s/m^3
t_b	1 s	R_p	$2.834{\cdot}10^{8}$ Pa·s/m^3

The governing flow equations and the structural equations, which are formulated, discretized and linearized in reference [5], are solved separately. Consequently, coupling iterations using the IQN-ILS algorithm [4] need to be performed between the flow equations and the structural equations to obtain the solution of the coupled problem. A cost function $y(\boldsymbol{x})$ is defined as the sum over all time steps and all segments of the squared difference between the radius in the simulation and in the measurement. This measurement, which would normally be obtained from a non-invasive medical imaging technique such as ultrasound, is mimicked by a simulation with the same model. It is then assumed that the parameter values in this “measurement simulation” have been forgotten and their values are calculated using the parameter identification. The vector $\boldsymbol{x}$ contains the k parameters which are defined in 37 and 38. The state vector $\boldsymbol{s}$ contains the radius in all segments and all time steps. The parameter identification can thus be reformulated as a minimization problem

$$\min_{\boldsymbol{x},\boldsymbol{s}} y(\boldsymbol{x},\boldsymbol{s}) \tag{39}$$

subject to the governing equations as constraints. As the state vector $\boldsymbol{s}$ depends on the parameters $\boldsymbol{x}$, the total derivative of the cost function $y(\boldsymbol{x},\boldsymbol{s}) = y(\boldsymbol{x},\boldsymbol{s}(\boldsymbol{x}))$ with respect to the parameters is obtained with the chain rule.

$$\frac{\mathrm{d}y}{\mathrm{d}\boldsymbol{x}} = \frac{\partial y}{\partial \boldsymbol{x}} + \frac{\partial y}{\partial \boldsymbol{s}}\frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}\boldsymbol{x}} \tag{40}$$

To avoid the direct calculation of $\mathrm{d}\boldsymbol{s}/\mathrm{d}\boldsymbol{x}$, the adjoint equations of this unsteady fluid-structure interaction problem are derived and solved, which involves backward time steps. In each of these steps, the adjoint flow equations and adjoint structural equations are coupled using the IQN-ILS algorithm [4], similarly to the forward equations.

The stiffness off each segment is identified by constructing a surrogate model for $y(\boldsymbol{x})$, followed by a search for its minimum and the corresponding values of $\boldsymbol{x}$.

C Surrogate model accuracy

Table 14: Prediction of Derivatives

Benchmark Function		CV Error	[10 - fold]	[$n_s = 100$	140(20D)]	
	Gaussian		Matérn $\frac{3}{2}$		Matérn $\frac{5}{2}$	
	OK	GEK	OK	GEK	OK	GEK
Ackley-3D	5.52e-01	3.18e-01	5.23e-01	3.83e-01	5.45e-01	3.53e-01
Hartmann-6D	2.88e-01	1.52e-01	2.93e-01	1.98e-01	2.93e-01	1.74e-01
Rosenbrock-8D	8.13e-01	4.64e-02	4.38e-01	2.08e-01	4.24e-01	1.31e-01
Rosenbrock-20D	1.63e+00	3.72e-01	1.63e+00	3.67e-01	1.63e+00	3.17e-01

Table 15: Efficiency of GEK and Matérn class of correlation functions (Prediction of Derivatives)

Benchmark Function		% of Improvement	[10 - fold]	[$n_s = 100/$	140(20D)]	
	Gaussian		Matérn $\frac{3}{2}$		Matérn $\frac{5}{2}$	
	OK	GEK	OK	GEK	OK	GEK
Ackley-3D	-	42%	5%	31%	1%	36%
Hartmann-6D	-	47%	-2%	31%	-2%	39%
Rosenbrock-8D	-	94%	46%	74%	48%	84%
Rosenbrock-20D	-	77%	0%	78%	0%	81%

Table 16: Efficiency of GEK (10D FSI Function)

n_s		Derivative Prediction	(Matérn $\frac{5}{2}$)	
	OK		GEK	
	NRMSE	% of improvement	NRMSE	% of improvement
20	5.46e-01	–	1.98e-01	64% better
40	5.46e-01	-7.3e-03% better	1.82e-01	67% better
60	5.46e-01	-5.0e-03% better	1.41e-01	74% better
80	5.46e-01	2.9e-03% better	1.31e-01	76% better
100	5.46e-01	1.0e-03% better	1.27e-01	78% better

Table 17: Prediction of Derivatives (FSI Functions)

FSI Function		CV Error	[10 - fold]	[$n_s = 100$]		
	Gaussian		Matérn $\frac{3}{2}$		Matérn $\frac{5}{2}$	
	OK	GEK	OK	GEK	OK	GEK
6D	3.67e-01	1.10e-01	3.97e-01	1.91e-01	3.75e-01	1.12e-01
7D	6.53e-01	1.72e-01	4.20e-01	2.47e-01	4.15e-01	1.65e-01
8D	9.37e-01	2.07e-01	5.48e-01	2.72e-01	5.30e-01	1.83e-01
9D	1.01e+00	2.31e-01	5.52e-01	3.00e-01	5.35e-01	2.13e-01
10D	1.06e+00	2.34e-01	5.59e-01	3.01e-01	5.36e-01	2.15e-01