Turning Black-Box Functions Into White Functions

Songqing Shan
Department of Mechanical and Manufacturing Engineering,
University of Manitoba,
Winnipeg, MB, R3T 5V6, Canada
e-mail: shans@cc.umanitoba.ca

G. Gary Wang[1]
School of Engineering Science,
Simon Fraser University,
Surrey, BC, V3T 0A3, Canada
e-mail: gary_wang@sfu.ca

A recently developed metamodel, radial basis function-based high-dimensional model representation (RBF-HDMR), shows promise as a metamodel for high-dimensional expensive black-box functions. This work extends the modeling capability of RBF-HDMR from the current second-order form to any higher order. More importantly, the modeling process "uncovers" black-box functions so that not only is a more accurate metamodel obtained, but also key information about the function can be gained and thus the black-box function can be turned "white." The key information that can be gained includes: (1) functional form, (2) (non)linearity with respect to each variable, and (3) variable correlations. The black-box "uncovering" process is based on identifying the existence of certain variable correlations through two derived theorems. The adaptive process of exploration and modeling reveals the black-box functions until all significant variable correlations are found. The black-box functional form is then represented by a structure matrix that can manifest all orders of correlated behavior of the variables. The resultant metamodel and its revealed inner structure lend themselves well to applications such as sensitivity analysis, decomposition, visualization, and optimization. The proposed approach is tested with theoretical and practical examples. The test results demonstrate the effectiveness and efficiency of the proposed approach.
[DOI: 10.1115/1.4002978]

Keywords: approximation, regression, interpolation, metamodel, response surface, prediction, RBF-HDMR, functional form, black-box function

1 Introduction

Metamodeling techniques find a wide range of uses in engineering. These uses include getting insight into a complex system and supporting simulation-based optimization. Metamodeling techniques involve sampling approaches, model selection, model fitting, and model validations [1–3]. From the sampling perspective, there are various sampling approaches: one-stage sampling (e.g., LATIN HYPERCUBE and orthogonal arrays), optimal sampling (e.g., D-optimal and G-optimal), one-stage-based optimal sampling (e.g., optimal LATIN HYPERCUBE, and optimal orthogonal-arrays-based LATIN HYPERCUBE) and sequential sampling [1–4]. For metamodel selection and fitting, there are parametric models (polynomial) and nonparametric models (radial basis function) [3]. For metamodel validation, there are various validating approaches and performance metrics such as relative error and R^2 [5]. There are multiple papers that reviewed the advancement of the metamodeling techniques. Chen et al. [6] summarized the pros and cons of the sampling methods and metamodels. Queipo et al. [7] reflected on the metamodeling techniques and optimization. Simpson et al. [8] reviewed the use of metamodeling techniques in multidisciplinary analysis and optimization. Wang and Shan [9] reviewed applications of metamodeling techniques in support of engineering optimization.

With the advancement of metamodeling techniques, two issues become prominent: (1) the "curse of dimensionality" and (2) how to gain more insight into a black box through metamodeling. The two issues interweave with each other. The "curse of dimensionality" indicates the metamodeling cost (the number of function evaluations) exponentially increases as the dimensionality of the black-box functions becomes larger. Koch et al. [10] presented the size problem of the black-box functions for multidisciplinary design optimization. Simpson et al. [11] pointed out that the high-dimensionality plagues metamodeling techniques. Shan and Wang [5] reviewed the relevant methodologies in solving high-dimensional expensive black-box (HEB) problems. Friedman and Stuetzle [12] developed projection pursuit regression. Friedman [13] proposed the multivariate adaptive regression spline (MARS) model. A family of high-dimensional model and representations (HDMRs) with distinct characters has since been developed for various purposes [14–19]. Recently Shan and Wang [20] proposed a RBF-HDMR model and its modeling algorithm to approximate a black-box function truncated after the second-order terms. For the second issue on gaining more insight into the black-box function, almost all the metamodels only provide the metamodel as a predictor and lack the capability to reveal the underlying functional form of the black-box function. This is particularly the case for commonly used metamodels such as Kriging, Support Vector Machine, and Radial Basis Function (RBF). Few papers have addressed this issue, especially from the metamodeling community. For instance, Booker [21] used functional ANOVA techniques in conjunction with a fitted Kriging model to disclose the main effects and correlation relationships in the response. Hooker [22] developed an ANOVA approach from statistics to discover additive structure in black-box functions.

This paper naturally advances the RBF-HDMR model to discover the intrinsic structure of a black-box function and enhance the accuracy of modeling by adaptively modeling higher-order terms beyond the current second-order form. For situations that the accuracy of the RBF-HDMR approximation provided by the first and second order is not sufficient, the developed approach explores high-order correlated terms and continues to model the residual terms. Thus, the accuracy of the RBF-HDMR can be further improved as far as the budget will allow. Moreover, the

[1]Corresponding author.
Contributed by the Design Automation Committee of ASME for publication in the JOURNAL OF MECHANICAL DESIGN. Manuscript received May 3, 2010; final manuscript received September 3, 2010; published online February 22, 2011. Assoc. Editor: Bernard Yannou.

modeling process can reveal key information about a black-box function such as the functional form, (non)linearity, and variable correlations.

Section 2 introduces the basics of RBF-HDMR. Section 3 describes the modeling process of the RBF-HDMR approach. Section 4 presents a structure matrix, a component correlation matrix, and theories that support identification of high-order component terms. Section 5 provides test results and discussion. The final remarks are in Sec. 6.

2 RBF-HDMR

A general form of an HDMR [16] is shown as

$$f(\boldsymbol{x}) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{1\le i<j\le d} f_{ij}(x_i,x_j) + \cdots + \sum_{1\le i_1<\cdots<i_l\le d} f_{i_1 i_2,\ldots,i_l}(x_{i_l},x_{i_2},\ldots,x_{i_l}) \cdots + f_{12,\ldots,d}(x_1,x_2,\ldots,x_d) \quad (1)$$

where f_0 is a constant representing the zeroth-order effect on $f(\boldsymbol{x})$; $f_i(x_i)$ is the effect of the variable x_i acting independently on the output $f(\boldsymbol{x})$ (the first-order effect) and may have an either linear or nonlinear dependence on x_i; $f_{ij}(x_i,x_j)$ describes the correlated contribution of the variables x_i and x_j on the output $f(\boldsymbol{x})$ (the second-order effect) after the individual influences of x_i and x_j are discounted, and $f_{ij}(x_i,x_j)$ could be linear or nonlinear as well. The subsequent terms reflect the effects of increasing numbers of correlated variables acting together on the output $f(\boldsymbol{x})$. The last term $f_{12\cdots d}(x_1,x_2,\ldots,x_d)$ represents the residual influence of all the variables locked together on the output $f(\boldsymbol{x})$ after all of the lower-order correlation and individual influence of each involved x_i ($i=1,\ldots,d$) have been discounted.

In order to compute component functions in Eq. (1), the simplest and most efficient type, Cut-HDMR [16], is explained here. For a chosen cutting center point x_0, component functions of the Cut-HDMR are defined as

$$f_0 = f(x_0) \quad (2)$$

$$f_i(x_i) = f(x_i,x_0^i) - f_0 \quad (3)$$

$$f_{ij}(x_i,x_j) = f(x_i,x_j,x_0^{ij}) - f_i(x_i) - f_j(x_j) - f_0 \quad (4)$$

$$f_{ijk}(x_i,x_j,x_k) = f(x_i,x_j,x_k,x_0^{ijk}) - f_{ij}(x_i,x_j) - f_{ik}(x_i,x_k) - f_{jk}(x_j,x_k) - f_i(x_i) - f_j(x_j) - f_k(x_k) - f_0 \quad (5)$$

$$f_{12\cdots d}(x_1,x_2,\ldots,x_d) = f(\boldsymbol{x}) - f_0 - \sum\nolimits_i f_i(x_i) - \sum\nolimits_{ij} f_{ij}(x_i,x_j) - \cdots \quad (6)$$

where x_0^i, x_0^{ij}, and x_0^{ijk} are respectively x_0 without elements x_i; x_i,x_j; and x_i,x_j,x_k. For the convenience of later discussions, the points x_0, $(x_i,x_0^i)=[x_{1_0},x_{2_0},\ldots,x_i,\ldots,x_{d_0}]^T$, $(x_i,x_j,x_0^{ij})=[x_{1_0},x_{2_0},\ldots,x_i,\ldots,x_j,\ldots,x_{d_0}]^T$, are, respectively, called the zeroth-order, first-order, second-order model-construction point(s), and so on. Accordingly, $f(x_0)$ is the value of $f(\boldsymbol{x})$ at x_0; $f(x_i,x_0^i)$ is the model output at point (x_i,x_0^i). The Cut-HDMR, in its original form, only provides a lookup table for data interpolation; there is no explicit expression for component functions. It also does not have a sampling approach to support HDMR construction.

The recently developed RBF-HDMR [20] uses a sum of a thin plate spline function (the first term) and a linear polynomial $P(\boldsymbol{x})$ (the second term) to approximate each component function in Eqs. (3)–(6).

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i |x-x_i|^2 \log|x-x_i| + P(\boldsymbol{x})$$

$$\sum_{i=1}^{n} \beta_i \boldsymbol{p}(x) = 0$$

$$P(\boldsymbol{x}) = \boldsymbol{p}\boldsymbol{\alpha} = [p_1,p_2,\ldots,p_q][\alpha_1,\alpha_2,\ldots,\alpha_q]^T \quad (7)$$

where x_i are the vectors of n evaluated sample points; the coefficients $\boldsymbol{\beta}=[\beta_1,\beta_2,\ldots,\beta_n]$ and $\boldsymbol{\alpha}$ are parameters to be found. $P(\boldsymbol{x})$ is a polynomial function, where $\boldsymbol{p}$ consists of a vector of basis polynomials. In this work, $\boldsymbol{p}$ is chosen to be $(1,x_1,\ldots,x_d)$ including only linear variable terms and therefore $q=d+1$; The side condition $\sum_{i=0}^{n}\beta_i \boldsymbol{p}(x)=0$ is imposed on the coefficients $\boldsymbol{\beta}$ to improve an underdetermined system, that is, the singularity of distance matrix $\boldsymbol{A}$ with $A_{ij}=|x_i-x_j|^2\log|x_i-x_j|$, $i,j=1,\ldots,n$. RBF is a simple interpolative function and found to provide a good approximation for arbitrary systems [6].

For the ease of description, we use a linear RBF as a substitute for Eq. (7) without losing generality. Therefore, a general RBF-HDMR model is written as

$$f(\boldsymbol{x}) \cong f_0 + \sum_{i=1}^{d}\sum_{k=1}^{m_i} \alpha_{i_k}|(x_i,x_0^i) - (x_{i_k},x_0^i)| + \sum_{1\le i<j\le d}\sum_{k=1}^{m_{ij}} \alpha_{ij_k}|(x_i,x_j,x_0^{ij}) - (x_{i_k},x_{j_k},x_0^{ij})| + \cdots + \sum_{k=1}^{m_{12,\ldots,d}} \alpha_{12,\ldots,d_k}|x-x_k| \quad (8)$$

where $|.|$ denotes a p-norm distance; $\alpha_{i_k},\alpha_{ij_k},\ldots,\alpha_{12,\ldots,d_k}$ are respectively the coefficients of the expression; $(x_{i_k},x_0^i),(x_{i_k},x_{j_k},x_0^{ij}),\ldots,x_k$ are the sampled points; $m_i,m_{ij},\ldots,m_{12,\ldots,d}$ are the number of sampled points for each term; the component $\sum_{k=1}^{m_i}\alpha_{i_k}|(x_i,x_0^i)-(x_{i_k},x_0^i)|$ is a function of only the ith input variable x_i, and explains the effect of the ith input variable x_i independently acting on the output function $f(\boldsymbol{x})$; and the component $\sum_{k=1}^{m_{ij}}\alpha_{ij_k}|(x_i,x_j,x_0^{ij})-(x_{i_k},x_{j_k},x_0^{ij})|$ denotes the correlated contribution of the variables x_i and x_j on the output $f(\boldsymbol{x})$ after the individual influences of x_i and x_j are discounted, and so on.

For a black-box function with d dimensionality, the number of all possible existing components can be expressed as

$$N = \sum_{i=0}^{d} \frac{d!}{(d-i)!i!} = 2^d \quad (9)$$

It can be seen that N increases dramatically as the dimensionality d rises. This challenges both the identification of the functional form and the modeling accuracy if higher-order correlated terms exist in the black-box functions. The RBF-HDMR has an attractive feature that it interpolates all of the prescribed points used for constructing all component functions. The prescribed points are defined as follows. For the constant component, the model-construction point is x_0; for the first-order components, the model-construction points are x_0 and (x_{i_k},x_0^i); and for the second-order components, its model-construction points are x_0, (x_{i_k},x_0^i), (x_{j_k},x_0^i), and $(x_{i_k},x_{j_k},x_0^{ij})$.

3 RBF-HDMR Modeling Process

Based on the argument that most well-defined physical systems involve relatively low-order correlations of the input variables [16,20], the RBF-HDMR modeling process up to the second-order was described as follows [20]:

(1) Randomly choose a point $x_0=[x_{1_0},x_{2_0},\ldots,x_{d_0}]^T$ in the modeling domain. Evaluating $f(\boldsymbol{x})$ at x_0, we then have f_0.
(2) Sample the first-order component functions $f_i(x_i)$

$=f([x_{1_0},x_{2_0},\ldots,x_i,\ldots,x_{d_0}]^T)-f_0$ in the close neighborhood of the two ends of x_i (lower and upper limits) while fixing the rest of x_j ($j\neq i$) components at x_0. Evaluating these two end points, gave us the left point value $f_{iL}(x_i)$ $=f([x_{1_0},x_{2_0},\ldots,x_{i_L},\ldots,x_{d_0}]^T)-f_0$ and similarly, the right point value $f_{iR}(x_i)$. Model the component function as $\hat{f}_i(x_i)$ with a 1D RBF model for each variable x_i using Eq. (7).

(3) Check the linearity of $f_i(x_i)$. If the approximation model $\hat{f}_i(x_i)$ goes through the center point x_0, then $f_i(x_i)$ is considered linear. In this case, modeling for this component terminates. Otherwise, use the center point x_0 and the two end points to reconstruct $\hat{f}_i(x_i)$. Then, a random value along x_i is generated and combined with the rest of the x_j ($j\neq i$) components at x_0 to form a new point to test $\hat{f}_i(x_i)$. If $\hat{f}_i(x_i)$ is not sufficiently accurate (the relative error is larger than a given criterion, for instance, 0.01%), the test point and all the evaluated points will be used to reconstruct $\hat{f}_i(x_i)$. This sampling-remodeling process iterates until convergence. This process captures the nonlinearity of the component function with one sample point at a time. Repeat Step 3 for all of the first-order component functions to construct the first-order terms of RBF-HDMR model.

(4) Form a new point, $(x_1,x_2,\ldots x_i,\ldots x_d)_k$ $=[x_{1_k},x_{2_k},\ldots,x_{i_k},\ldots,x_{j_k},\ldots,x_{d_k}]^T$, $k\neq 0$ by randomly combining the sampled value x_i in the first-order component construction for each input variable (that is, x_i, $i=1,\ldots,d$ in the evaluated (x_i,x_0^i), respectively). This new point is then evaluated by expensive simulation, as well as by the first-order RBF-HDMR model. If the two function values are sufficiently close (the relative error is less than a small value, for example, 0.01%), it indicates that no higher-order terms exist in the underlying function, and the modeling process terminates. Otherwise, go to Step 5.

(5) Use the values of x_i and x_j, $i\neq j$ that exist in the points evaluated thus far $(x_i,x_0^i)=[x_{1_0},x_{2_0},\ldots,x_i,\ldots,x_{d_0}]^T$ and $(x_j,x_0^j)=[x_{1_0},x_{2_0},\ldots,x_j,\ldots,x_{d_0}]^T$ to create new points of the form $(x_i,x_j,x_0^{ij})=[x_{1_0},x_{2_0},\ldots,x_i,\ldots,x_j,\ldots,x_{d_0}]^T$. Randomly select one point from these new points to test the first-order RBF-HDMR model. If the model passes through the new point, it indicates that x_i and x_j are not correlated, and the process continues with the next pair of input variables. This is to save the cost of modeling nonexistent or insignificant correlations. Otherwise, use this new point and the evaluated points (x_i,x_0^i) and (x_i,x_0^j) to construct the second-order component function, $\hat{f}_{ij}(x_i,x_j)$. This sampling-remodeling process iterates for all possible two-variable correlations until convergence (the relative error is less than 0.01%). Step 5 is repeated for all pairs of input variables.

Theoretically, step 5 applies to all higher-order terms in the RBF-HDMR model in a similar manner. However, given the exponentially growing number of terms as shown in Eq. (9), even if only one extra point is needed to test whether or not a higher-order correlation exists (such as that in step 5 for bivariate correlation), the number of sample points needed would increase exponentially. Therefore, this work first introduces some theorems and then uses the theorems to guide the modeling of higher-order component functions.

Figure 1 shows a simplified flow of the RBF-HDMR modeling process. The step Refine Model means increasing the number of samples to improve the accuracy of the model without changing the functional form of the model; the step Update Functional Form adds the correlated component terms to the RBF-HMDR model if the term exists. When the desired modeling accuracy is reached, the modeling process terminates. The last step for higher-order components is discussed in the next section.

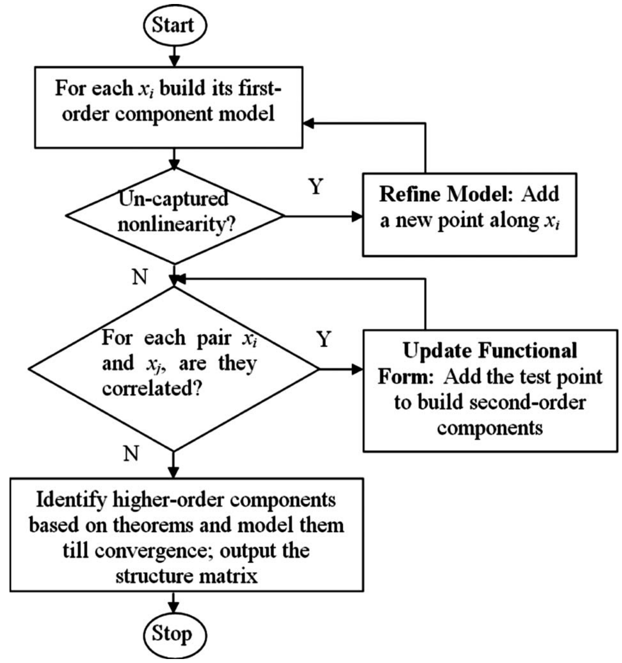


Fig. 1 A simplified flow of RBF-HDMR metamodeling

4 Identification of Functional Form

First, this section first defines two matrices that support the identification of multivariate correlated terms and the functional form. Then, theorems for identification are introduced, which form the basis for efficiently modeling higher-order components of the RBF-HMDR and for uncovering the functional form of the black-box function.

4.1 Structure Matrix. A structure matrix (SM) is defined to capture the inner structure of the resultant RBF-HDMR of a black-box function as

$$\mathrm{SM}_{d\times n}=\begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 & 1 & \ldots & 0 & \ldots & 0 & \ldots & 1\\ 0 & 0 & 1 & 0 & \ldots & 0 & 1 & \ldots & 0 & \ldots & 0 & \ldots & 1\\ 0 & 0 & 0 & 1 & \ldots & 0 & 0 & \ldots & 0 & \ldots & 0 & \ldots & 1\\ & & & & & \ldots & & & & & & & \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & 1 & \ldots & 1\\ 0 & 0 & 0 & 0 & \ldots & 1 & 0 & \ldots & 0 & \ldots & 1 & \ldots & 1 \end{bmatrix} \tag{10}$$

where d is the dimensionality of the input variable vector x; n denotes the number of to-be-decided component terms. Each row corresponds to a variable x_i. Each column corresponds to one of the component terms in the RBF-HDMR. Each element in the structure matrix is assigned as "0" or "1;" "0" means that the variable is sampled at x_{i_0}; "1" means that the variable is sampled at non-x_{i_0} locations. For example, the column $[0_1,0_2,\ldots,0_d]^T$ denotes the constant component term f_0; $[0_1,0_2,\ldots,1_0,\ldots,0_d]^T$ represents the first-order component term $f_i(x_i)$; $[0_1,0_2,\ldots,1_i,\ldots,1_j,\ldots,0_d]^T$ indicates the existence of the second-order component term $f_{ij}(x_i,x_j)$, and the last column indicates the existence of a d-variate correlation component terms $f_{12,\ldots,d}(x_1,x_2,\ldots,x_d)$.

The structure matrix is employed to index the corresponding component term and is created in tandem with the RBF-HDMR modeling process. Since each column in the structure matrix is associated with a unique component term and the maximum number of terms is 2^d (Eq. (9)), a structure matrix could theoretically represent a maximum of 2^d terms [16]. Many component terms may not exist in the black-box function $f(\boldsymbol{x})$, and some others have negligible contributions to $f(\boldsymbol{x})$. These terms' corresponding columns are eliminated from the structure matrix. For example, if the column $[0_1, 0_2, \dots, 1_i, \dots, 1_j, \dots, 0_d]^T$ does not exist in the structure matrix, it means $f_{ij}(x_i, x_j)$ does not exist or is negligible. For descriptive convenience, a nonexistent or negligible term is referred to in the rest of the paper as an insignificant term; otherwise, it is a significant term. The final output of the structure matrix thus depends on the intrinsic characteristics of the black-box function, and the structure matrix in return explicitly reveals the inner functional form of the black-box function. Each column in the structure matrix represents one term in the final RBF-HDMR. For each element x_i, a "1" in a column means that the variable exists in the corresponding component term.

4.2 Component Correlation Matrix. Given the fact that HDMR is built on a hierarchy of orthogonal component functions with increasing dimensionality, we can further explore the variable relationships in the context of component functions. For instance, as one understands from Eqs. (2)–(6), $f_{12}(x_1, x_2)$ does not simply capture the term $x_1 x_2$, but rather the residual effect $(f(x_1, x_2, \boldsymbol{x}_0^{12}) - f_1(x_1) - f_2(x_2) - f_0)$ of $x_1 x_2$. In other words, the algebraic term $x_1 x_2$ is expressed by f_0, $f_1(x_1)$, $f_2(x_2)$ and $f_{12}(x_1, x_2)$ altogether. The effect of the term $x_1 x_2$ may be well captured by the first-order components, and thus there is no need to model the second-order term $f_{12}(x_1, x_2)$; otherwise, $f_{12}(x_1, x_2)$ needs to be added to accurately model the $x_1 x_2$ term. Whether or not $f_{12}(x_1, x_2)$ is significant, it helps us to define the variable correlation $x_1 x_2$ from the perspective of HDMR. This point separates our variable correlation matrix, to be defined below, from its conventional meaning. To distinguish the difference, we define the component correlation matrix as follows.

Considering d variables (that is, $\boldsymbol{x} = [x_1, x_2, \dots, x_d]^T$), both the row list and the column list in a component correlation matrix denote the same set of input variables. The matrix entry indicates whether the ith input variable in row and the jth input variable in column defines a component term $f_{ij}(x_i, x_j)$. A component correlation matrix (CCM) is

$$\mathrm{CCM}_{d\times d} = [m_{ij}] \quad (i = 1, 2, \dots, d;\ j = 1, 2, \dots, d) \tag{11}$$

where $m_{ij} = 1$, if $f_{ij}(x_i, x_j)$ exists in an HDMR formula for a particular problem; otherwise, $m_{ij} = 0$. For example, a component correlation matrix for a function of only first-order components is a diagonal matrix of 1's; the component correlation matrix of all significant bivariate component terms is a square matrix with all 1's, as shown in Fig. 2. Similar to a conventional correlation matrix, a CCM is symmetric.

A CCM can be automatically generated after completely modeling RBF-HDMR's second-order terms because the modeling process adaptively identifies such relationships. The 0's and 1's scatter in the CCM, depending on the characteristics of the underlying black-box function. A CCM can be reorganized by changing the order of rows and columns to exhibit patterns of correlations or extract a part of the rows and columns to form subcorrelation matrices. A CCM captures all of the bivariate component terms and leads to the identification of more-than-two-variable component terms, which will be discussed in Sec. 4.3.

4.3 Correlation Identification for Higher-Order Component Modeling. A CCM matrix shows second-order component terms between variable pairs. How can we identify higher-order component terms involving three or more variables without incurring extra sampling costs? Let us take a t-variable subset $(3 \le t \le d)$ from a CCM to form a new submatrix. Such matrices include two types, a $t \times t$ matrix with all 1's and a $t \times t$ matrix with at least one 0.

$\boldsymbol{x}$	x_1	x_2	$\cdots$	x_i	$\cdots$	x_d
x_1	1	1	$\cdots$	1	1	1
x_2	1	1	$\cdots$	1	1	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	1	1	$\cdots$	1	1	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_d	1	1	$\cdots$	1	1	1

Fig. 2 An example of component correlation matrix indicating a function having all significant bivariate terms

THEOREM 1. *The necessary condition of a t-variable* $(t \ge 3)$ *component term existing in an HDMR formulation for a black-box function is that the t-variable submatrix of the CCM is a* $t \times t$ *matrix with all 1's and all of its component terms involving* $(t-1)$ *variables exist in the HDMR model.*

Proof. Assuming t-variables $x_{i_1}, \dots, x_{i_t}$, if one possible component term of a subset of $(t-1)$ variables, $x_{i_1}, \dots, x_{i_{t-1}}$, $f_{i_1,\dots,i_{t-1}}$, does not exist, it means that the contribution of the subset $x_{i_1}, \dots, x_{i_{t-1}}$ is not significant after all the lower-order effects are modeled. Therefore, the higher-order component $f_{i_1,\dots,i_t}$ would not be significant either. Similarly if $f_{i_1,\dots,i_t}$ exists, it means that all the lower-order components for all the t-variables should exist since $f_{i_1,\dots,i_t}$ is computed from all its related lower-order components (see Eqs. (5) and (6)). Therefore, all entries in the $t \times t$ matrix should be 1's. If there exists an entry of "0" between two variables x_{i_l} and x_{i_m}, it means that $f_{i_l i_m}$ does not exist or is not significant, and therefore any higher-order components involving x_{i_l}, x_{i_m} would not be significant either. Proof completed.

Theorem 1 can be used to explore higher-order component terms in an RBF-HDMR. If the necessary condition is not met, then the corresponding t-variable component term does not exist, and that term is skipped during modeling, thus an extra sample point is saved. It is to be noted if a submatrix of a t-variable CCM has all entry of 1's, one cannot sufficiently conclude that all of the third- and higher-order components exist. This is because a CCM only defines bivariate relations. Theorem 2 is therefore proposed to supplement Theorem 1.

THEOREM 2. *The sufficient condition of the existence of a t-variable* $(t \ge 3)$ *component term in an RBF-HDMR formulation for a black-box function is that the value of a new point (formed from the existing model-construction points' variable elements for up to the* $(t-1)$th *order component terms) is not accurately predicted by the RBF-HDMR model of* $(t-1)$th *order.*

Proof. Assume an RBF-HDMR model of $(t-1)$th order is built, that is,

$$\tilde{f}(\boldsymbol{x}) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i \le 1 < j \le d} f_{ij}(x_i, x_j) + \dots \sum_{1 \le i_1 < \dots < i_{t-1} \le d} f_{i_1 i_2, \dots, i_{(t-1)}}(x_{i_1}, x_{i_2}, \dots, x_{i_{(t-1)}}) \tag{12}$$

All the model-construction points up to the $(t-1)$th order include x_0, (x_i, x_0^i), $(x_i, x_j, \dots, x_0^{ij})$, and $(x_{i_1}, x_{i_2}, \dots, x_{i_{(t-1)}}, x_0^{i_1 i_2, \dots, i_{(t-1)}})$. One picks variable elements from these points to form a new point $(x_{i_1}, x_{i_2}, \dots, x_{i_t}, x_0^{i_1 i_2, \dots, i_t})$. If the current $(t-1)$th order RBF-HDMR

GIVEN: CCM obtained after modeling the 2^{nd} order component terms

FOR each t-variable sub-matrix $(t = 3, \cdots, d)$

IF there is one '0' entry, skip modeling the t-variable component term (Theorem 1)

ELSE

Construct a new point according to Theorem 2 and evaluate the point

IF model prediction is not accurate, then model the *t*-variable component term (Theorem 2)

ELSE skip modeling the *t*-variable component term (Loosened condition)

Update the structure matrix

END

Fig. 3 Process for high-order component identification

$$SM = \left[\begin{array}{c|cccccccccc|ccccccccccc|ccccc|c}
0 & 1&0&0&0&0&0&0&0&0&0 & 1&1&0&0&0&0&0&0&0&0&0 & 1&0&0&0&0 & 0\\
0 & 0&1&0&0&0&0&0&0&0&0 & 1&0&1&1&0&0&0&0&0&0&0 & 1&0&0&0&0 & 0\\
0 & 0&0&1&0&0&0&0&0&0&0 & 0&1&1&0&1&0&0&0&0&0&0 & 1&0&0&0&0 & 0\\
0 & 0&0&0&1&0&0&0&0&0&0 & 0&0&0&0&0&0&0&0&0&0&0 & 0&0&0&0&0 & 0\\
0 & 0&0&0&0&1&0&0&0&0&0 & 0&0&0&0&1&0&0&0&0&0&0 & 0&0&0&0&0 & 0\\
0 & 0&0&0&0&0&1&0&0&0&0 & 0&0&0&0&0&0&0&0&0&0&0 & 0&0&0&0&0 & 0\\
0 & 0&0&0&0&0&0&1&0&0&0 & 0&0&0&1&0&1&1&1&0&0&0 & 0&1&1&1&0 & 1\\
0 & 0&0&0&0&0&0&0&1&0&0 & 0&0&0&0&0&1&0&0&1&1&0 & 0&1&1&0&1 & 1\\
0 & 0&0&0&0&0&0&0&0&1&0 & 0&0&0&0&0&0&1&0&1&0&1 & 0&1&0&1&1 & 1\\
0 & 0&0&0&0&0&0&0&0&0&1 & 0&0&0&0&0&0&0&1&0&1&1 & 0&0&1&1&1 & 1
\end{array}\right]$$

Fig. 4 The structure matrix of the example

cannot accurately predict the function value at the new point, it means that the interaction of all the *t*-variables has not been captured by the $(t-1)$th order model and therefore the *t*-variable $(t \geq 3)$ correlated component term should exist. Proof completed.

Theorem 2, the sufficient condition, can be used for confirming the existence of the correlated component terms. However, what if the $(t-1)$th order model accurately predicts the function values of all possible new points? Strictly speaking, one cannot sufficiently conclude from Theorem 2 that there does not exist *t*th order or higher components. Assuming it is a rare occurrence for a black-box function having *t*th order or higher components uncaptured and yet the $(t-1)$th order model accurately predicts the function values of new points, the sufficient condition is loosened as follows for practical algorithm development.

4.3.1 Loosened Sufficient Condition. If an RBF-HDMR of $(t-1)$th order can exactly predicts the function value at the test point constructed from existing model points' variable elements for up to the $(t-1)$th order component terms, then it is deemed that there is no *t*-variable or higher-order component terms in the black-box function.

4.4 Modeling of High-Order Component Functions in RBF-HDMR. The theorems developed in Sec. 4.3 are employed to identify and model high-order component functions in an RBF-HDMR.

The general process follows that for the modeling of a second-order component, as described in Sec. 3. The main difference is on the identification of high-order component functions in order to avoid exponentially increasing sampling costs and to reveal the functional form of the black-box function. The logic for component identification is illustrated in Fig. 3.

The structure matrix of the black-box function, as shown in Eq. (10), is constructed and updated with the modeling process. Once the modeling is completed, its functional form is captured in the structure matrix. The following example explains the process in detail.

4.5 An Example. One example is used to illustrate the modeling and functional form identification process. This problem is modified from Ref. [22] and expressed as

$$f(x) = \pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1} x_4 + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7 + {x_6}^2 \tag{13}$$

The resultant structure matrix of this example is shown in Fig. 4. In the structure matrix, we shadow component terms from zeroth order to the highest order with gradually lighter colors. Also, we mark the independent component terms with rectangular boxes. Matching the structural matrix with steps in Sec. 3, the first colored block corresponds to step 1; the second colored block is implemented by steps 2 and 3; step 4 happens between the second colored block and the third colored block; step 5 fills the third colored block and generates a component correlation matrix; the last two colored blocks are implemented by means of the two derived theorems and the algorithm shown in Fig. 3. From the fourth and sixth rows, one can see that input variables x_4 and x_6 are independent variables, that is, having only first-order component terms. Observing the second rightmost block, one can see that variables x_1, x_2, and x_3 form a three-variable component term; observing the last block, variables x_7, x_8, x_9, x_{10} form a four-variable component term. The middle color block shows that two-variable correlation exists between variable pairs x_2 and x_7, as well as x_3 and x_5. It is to be noted that multiple variable component terms can be ignored if the terms are trivial to the output. From the final structure matrix, one can extract the following mathematical expression

$$f(x) = f_0 + g_4(x_4) + g_6(x_6) + g_{2,7}(x_2,x_7) + g_{3,5}(x_3,x_5) + g_{1,2,3}(x_1,x_2,x_3) + g_{7,8,9,10}(x_7,x_8,x_9,x_{10})$$

where $g_i(x_i) = f_i(x_i), \quad i = 4,6$

$$g_{2,7}(x_2,x_7) = f_2(x_2) + f_7(x_7) + f_{2,7}(x_2,x_7)$$

$$g_{3,5}(x_3,x_5) = f_3(x_3) + f_5(x_5) + f_{3,5}(x_3,x_5)$$

$$g_{1,2,3}(x_1,x_2,x_3) = \sum_{i=1}^{3} f_i(x_i) + \sum_{i=1,i<j}^{3} f_{ij}(x_i,x_j) + f_{1,2,3}(x_1,x_2,x_3)$$

$$\begin{aligned} g_{7,8,9,10}&(x_7,x_8,x_9,x_{10}) \\ &= \sum_{i=7}^{10} f_i(x_i) + \sum_{i=7,i<j}^{10} f_{ij}(x_i,x_j) + \sum_{i=7,i<j<k}^{10} f_{ijk}(x_i,x_j,x_k) \\ &\quad + f_{7,8,9,10}(x_7,x_8,x_9,x_{10}) \end{aligned} \tag{14}$$

Equation (14) corresponds to the structure depicted by Fig. 4. f_0 corresponds to the first column. The numerical models of all component functions have been obtained using the modeling process described in Sec. 3 and stored in the final model. The final model manifests the high-dimensional correlated behavior of variables. The linearity/nonlinearity information for each input variable is also saved in the final model and can be readily output.

The CCM corresponding to the SM in Fig. 4 is shown in Eq. (15)

Table 1 Modeling results of the example

Order	R^2	RAAE	RMAE	NoE id	NoE Model	NoE Total
First	0.3809	0.3331	10.399	0	137	137
Second	0.9386	0.1150	5.1044	35	238	273
Third	0.9182	0.1372	2.4582	35	358	393
Fourth	0.9187	0.1361	2.4580	35	406	441

$$\mathrm{CCM} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{15}$$

From this CCM, one can roughly see the correlation among the variables in Eq. (14). The top left corner 3×3 submatrix indicates that x_1, x_2, and x_3 may be correlated; however, it needs to be judged by the structure matrix. Similarly, the bottom right corner 4×4 submatrix indicates that x_7, x_8, x_9, and x_{10} may be correlated. Both the fourth row and fourth column have only one "1" element at the diagonal position, which shows that x_4 is only in first-order component term $f_4(x_4)$ and not in higher-order terms. The same is true for x_6. Variables x_2 and x_7 correlate strongly and $f_{2,7}(x_2,x_7)$ must be modeled; this is also true for x_3 and x_5.

The modeling result is given in Table 1, where NoE accumulates the number of function evaluations from lower to higher order; for example, it requires a total of 441 points to model the function up to the fourth order, 393 points to the third order, and so on. "id" means accumulated NoE spent on identification of the functional form, which is used for modeling if the term exists, and 35 sampling points are generated in the second order for this purpose. The column "model" indicates the NoE used for modeling. R^2, relative average absolute error (RAAE), and relative maximum absolute error (RMAE) are model performance metrics, which will be introduced in the next section. Table 1 shows that the second RBF-HDMR models the underlying function well, and usually no more modeling effort is needed. However, one can see that the performance metrics such as R^2 and RAAE became worse from the second to the third order. This phenomenon indicates "over fitting," this is, the gain from modeling higher-order terms is less than the error brought from the modeling process. Over fitting is one of the common issues in metamodeling techniques. Wang et al. [23] discussed over fitting in the RBF metamodel. Tecko et al. [24] presented a comparison of over fitting and over training in artificial neural networks. In the artificial neural network community, some additional techniques such as early stopping and cross-validation are used to avoid over fitting. Due to limited space, a discussion on over fitting is not extended here.

5 Test Examples

5.1 Problem Description. To test the effectiveness and efficiency of the proposed approach, 15 test problems are selected based on the criteria: (1) the number of variables ≥ 10, (2) high nonlinearity of the performance behavior, and (3) multiple variables are correlated. The criteria are chosen to expose the challenges of metamodeling HEB problems. Scalable problems with different dimensionality are treated as one problem. In total, 15 problems that satisfy the criteria are found in the book by Schittkowski [25], which offers 188 problems for testing nonlinear optimization algorithms and a few of them for testing data fitting algorithms. Most of these problems have some application background. Fifteen problems that satisfy our criteria are listed in the Appendix. Among these problems, the first ten are classified by Schittkowski as "theoretical" problems denoted by "T," and the remaining five problems as "practical" problems represented by "P." Detailed backgrounds of these practical problems are omitted due to limited space. "Order" stands for the highest order, that is analyzed by the RBF-HDMR. The modeling accuracy is evaluated by four performance metrics, which are introduced in the next section.

5.2 Performance Metrics

(1) R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{m}[f(x_i) - \hat{f}(x_i)]^2}{\sum_{i=1}^{m}[f(x_i) - \bar{f}(x_i)]^2} \tag{16}$$

where $\bar{f}(x_i)$ denotes the mean of the function on the m sampling points. These metrics indicate the overall accuracy of the approximation model. The closer the value of R^2 approaches 1, the more accurate the approximation model is. Note R^2 in this work is computed on 10,000 new test points for each problem, rather than on the modeling points. The same is true for the next two metrics.

(2) RAAE

$$\mathrm{RAAE} = \frac{\sum_{i=1}^{m} |f(x_i) - \hat{f}(x_i)|}{m * \mathrm{STD}} \tag{17}$$

where STD stands for standard deviation. Like R^2, this metric shows the overall accuracy of an approximation model. As the value of RAAE approaches zero, the approximation model becomes more accurate.

(3) RMAE

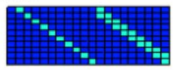
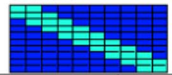
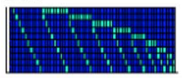
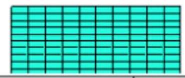
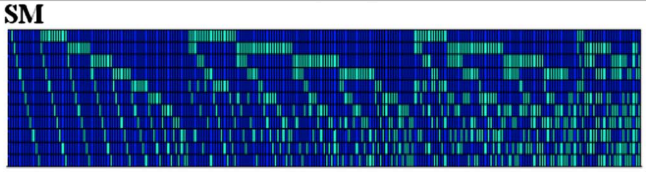
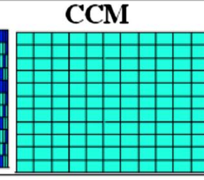
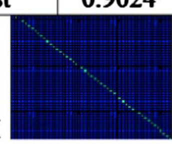
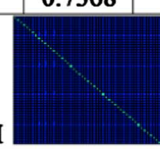
$$\mathrm{RMAE} = \frac{\max(|f(x_1) - \hat{f}(x_1)|, |f(x_2) - \hat{f}(x_2)|, \ldots, |f(x_m) - \hat{f}(x_m)|)}{\mathrm{STD}} \tag{18}$$

This is a local metric. An RMAE describes the error in a subregion of the design space. Therefore, a small value of RMAE is preferred.

5.3 Test Results. Expressions for the 15 problems are listed in Appendix. Table 2 shows the results of 14 test examples except for problem 12. Problem 12 is discussed in Sec. 5.4. The results represent the average of 10 independent runs. It can be seen that RBF-HDMR accurately models 14 problems out of 15.

In Table 2, problems 1, 2, 11, and 15 are chosen for detailed report; other results are also in Table 2 for brevity. For problems 1, 2, 11, and 15 in Table 2, the two matrices for each problem are shown, along with their mathematical function descriptions as those in Eq. (14). For problem 1, bivariate correlations exist, which are clearly shown with brighter colors in the two matrices. The RBF-HDMR model also reached high accuracy when including up to the second-order components. Problem 2 has high-order multivariate correlations, but these terms have small influence and can be neglected. Problem 11 has up to sixth-order correlations. The fourth-order RBF-HDMR model reaches an R^2 value of 0.938 and the cost grows significantly due to strong variable correlations as the model moves one order higher. Problem 15 has 50 variables, but its internal structure is very simple and there is no strong correlation between variables. Thus the modeling cost is low. Comparing problems 1 and 2, both have $d=10$, the NoE is

Table 2 Test results of examples

Function		R Square	RAAE	RMAE	NoE id	NoE model	NoE total
1 d=10	First	0.8683	0.2770	1.8064	0	51	51
	Second	0.9900	0.0804	0.3457	37	159	196
	SM 		CCM 		$f(\boldsymbol{x}) = c + \sum_{i=1}^{9} g_{i,i+1}(x_i, x_{i+1})$ No higher than 2nd order		
2 d=10	First	0.2756	0.4027	11.0485	0	51	51
	Second	0.9713	0.1025	2.7502	1	583	584
	SM 		CCM 		$f(\boldsymbol{x}) = c + \sum_{1 \le i < j \le 10} g_{i,j}(x_i, x_j)$ Higher-order negligible		
11 d=11	First	-0.3517	0.9050	5.6850	0	56	56
	Second	0.5412	0.4508	4.7432	1	579	580
	Third	0.8946	0.2368	2.0660	82	1251	1333
	Fourth	0.9368	0.1835	1.6048	82	2211	2293
	Only output first four orders						
	SM 					CCM 	
	Only write out 3 highest correlated terms $f(\boldsymbol{x}) = c + \cdots + g_{2,3,6,7,9,10}(x_2, x_3, x_6, x_7, x_9, x_{10}) + g_{2,4,6,8,9,11}(x_2, x_4, x_6, x_8, x_9, x_{11}) + f_{3,4,7,8,10,11}(x_3, x_4, x_7, x_8, x_{10}, x_{11});$ No higher than 6th order						
15 d=50	First	0.9024	0.2986	0.7368	0	251	251
	 SM		 CCM		$f(\boldsymbol{x}) = c + \sum_{i=1}^{50} g_i(x_i)$ No higher than 1st order		
3 d=10	First	0.8622	0.2888	1.6642	0	101	101
	Second	0.9438	0.2027	0.8231	136	209	345
4 d=20	First	0.8918	0.2555	1.3816	0	101	101
	Second	0.9471	0.2008	0.6896	166	265	431
5 d=20	First	−0.1862	0.8074	4.9560	0	101	101
	Second	0.8778	0.2663	1.5423	161	341	502
6 d=30	First	0.4029	0.5293	5.3301	0	151	151
	Second	0.8715	0.2161	4.6852	1	3631	3632
7 d=30	First	−2.6392	1.7225	6.869286	0	151	151
	Second	0.7746	0.4213	1.6282	1	5367	5368
8 d=31	First	0.5126	0.5494	3.1200	0	151	151
	Second	0.9506	0.1789	0.7747	407	499	906
9 d=20	First	0.4968	0.5585	4.2935	0	101	101
	Second	0.9906	0.0852	0.4121	172	265	437
10 d=20	First	0.9206	0.1936	3.1821	0	99	99
	Second	0.9971	0.0373	1.0299	1	1642	1643
13 d=14	First	0.9997	0.0135	0.0716	0	71	71
14 d=30	First	0.9999	0.0075	0.0168	0	151	151

196 and 584, respectively for the second-order model. As one can see that the cost for problem 2 is significantly higher. Problem 2 consists of multiple correlated second-order and third-order terms, which are evident from its SM and CCM. Its CCM has all elements of 1. The SM also appears more complicated than that of problem 1. Then for problem 11 with eleven variables, the cost for the second order is similar to that of problem 2 but doubles each time the order increases by 1.

Problem 11 has even more complex structure than problem 2 with multiple correlated high-order terms until the sixth order,

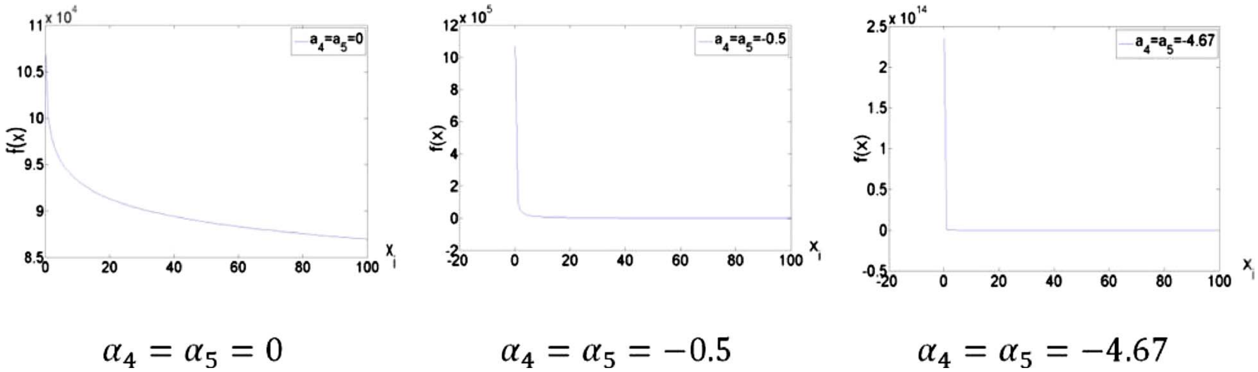


Fig. 5 Deterioration of $f(x)$ when decreasing coefficients α_4 and α_5

which explains its high modeling cost. For problem 15, a first-order problem, the NoE is only 251 even with $d=50$. Table 2 shows the test results for other functions. These results show the effectiveness and efficiency for a wide range of problems. In addition, this work did not choose to compare the RBF-HDMR with other metamodeling techniques. Interested readers can refer to Ref. [20] for such comparison.

5.4 Discussion. The test problem 12 is expressed as

$$f(\boldsymbol{x}) = 100000\prod_{i=1}^{11} x_i^{a_i}, \quad 0.1 \le x_i \le 100, \quad i=1,\ldots,11$$

$$\boldsymbol{\alpha} = (-0.00133172, -0.002270927, -0.00248546, -4.67, -4.671973, -0.00814, -0.008092, -0.005, -0.000909, -0.00088, -0.00119) \tag{19}$$

In this problem, α_i is employed for specifying the role of the corresponding variable x_i. α_i significantly affects the final output and the modeling results. This effect can be seen in Fig. 5. Figure 5 plots the $f(\boldsymbol{x})$ (vertical axis) with respect to x_i (all x_i are held equal to each other). As α_4 and α_5 get smaller, the output curve becomes extremely steep. When they reach a value at (−4.67, −4.67), the output forms a right angle with the x-axis and the modeling error is prohibitive. Table 3 shows the modeling results. The structure matrices and correlation matrices are given in Fig. 6. It is very interesting to see as the coefficients decrease from 0 to −0.5, the SM and the CCM change from a simple structure to complex ones with multivariate correlations. It means as α_i decreases, high-order variable correlations becomes stronger and finally dominates $f(\boldsymbol{x})$.

Table 3 The results of example 12

α_4	α_5	R^2	RAAE	RMAE	NoE
0	0	0.9914	0.0406	0.9234	126
−0.5	−0.5	0.6791	0.1438	20.1244	144
−4.67	−4.67	0.0000	0.0114	99.7205	146

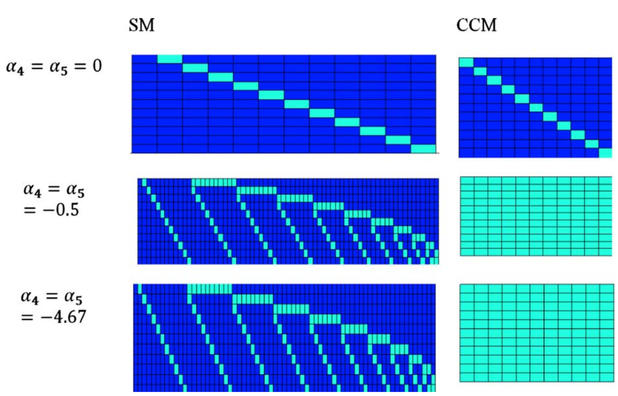


Fig. 6 Structure matrices and correlation matrices of problem 12

6 Final Remarks

This work extends the recently developed RBF-HDMR method to model higher than second-order component functions, based on which a black-box function can be "uncovered." Key information about a black-box function such as functional form, variable (non)linearity, and variable correlations can be obtained through the modeling process. SM is developed to present the functional form of the black-box problems. CCM is defined as to describe correlation relationship among the variables. Note that the SM and the CCM depend on the characteristics of HEB problems and do not dictate the exclusive association with RBF-HDMR. First, if other metamodels are used to model the component terms of HDMR, the SM and CCM remain exactly the same as in the context of RBF-HDMR. Second, even if HDMR is not used, for instance, for a second-order polynomial response surface, its constant, first-order and second-order terms correspond to the first few columns of SM as defined in Eq. (10); CCM can represent the correlation among variables according to the coefficients. However, the full second-order response surface is a parametric model whose functional form is postulated. Therefore, the high-order terms may be significant but ignored once the second-order polynomial model is chosen. Although the SM and the CCM are applicable for other metamodel techniques, the challenge lies in how to uncover the functional form and fill the matrices. In this work, two theorems are developed to support the efficient identification of high-order correlation terms in the context of HDMR. Multiple test examples show the effectiveness and efficiency of the proposed approach. Future work will extend the methodology to support ANOVA analysis, direct problem decomposition, and design optimization.

Acknowledgment

Funding supports from Canada Graduate Scholarships (CGS) and Natural Science and Engineering Research Council (NSERC) of Canada are gratefully acknowledged.

Appendix

No.	Function	Variable ranges	Class	Order
1	$f(\boldsymbol{x})=(x_1-1)^2+(x_{10}-1)^2+10\Sigma_{i=1}^{9}(10-i)(x_i^2-x_{i+1})^2$	$-3\le x_i\le 2$, $i=1,\ldots,10$	T	2
2	$f(\boldsymbol{x})=[\Sigma_{i=1}^{10} i^3(x_i-1)^2]^3$	$-3\le x_i\le 3$, $i=1,\ldots,10$	T	3
3	$f(\boldsymbol{x})=\Sigma_{i=1}^{10}[100(x_i-x_{i+10})^2+(x_i-1)^2]$	$-3\le x_i\le 5$, $i=1,\ldots,10$	T	2
4	$f(\boldsymbol{x})=\Sigma_{1}^{5}[100(x_i^2+x_{i+5})^2+(x_i-1)^2+90(x_{i+10}^2+x_{i+15})^2+(x_{i+10}-1)^2+10.1[(x_{i+5}-1)^2+(x_{i+15}-1)^2]+19.8(x_{i+5}-1)\times(x_{i+15}-1)]$	$-3\le x_i\le 5$, $i=1,\ldots,5$	T	2
5	$f(\boldsymbol{x})=\Sigma_{1}^{5}[(x_i+10x_{i+5})^2+5(x_{i+10}-x_{i+15})^2+(x_{i+5}-2x_{i+10})^4+10(x_i-x_{i+15})^4]$	$-2\le x_i\le 5$, $i=1,\ldots,5$	T	2
6	$f(\boldsymbol{x})=1-\exp[-1/60\Sigma_{i}^{30}x_i^2]$	$0\le x_i\le 3.5$, $i=1,\ldots,30$	T	30
7	$f(\boldsymbol{x})=(\boldsymbol{x}^T A\boldsymbol{x})^2,\quad A=\mathrm{diag}(1,2,3,\ldots,30)$	$-2\le x_i\le 3$, $i=1,\ldots,30$	T	2
8	$f(\boldsymbol{x})=\Sigma_{i=1}^{29}[100(x_{i+1}-x_i^2)^2+(1-x_i)^2]$	$-2\le x_i\le 2$, $i=1,\ldots,30$	T	2
9	$f(\boldsymbol{x})=\boldsymbol{x}^T A\boldsymbol{x}-2x_1,\quad A=\begin{bmatrix} 1 & -1 & & & & & 0 \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ldots & \ldots & \ldots & & 0 \\ & & & \ldots & \ldots & \ldots & \\ & & & -1 & 2 & -1 & \\ 0 & & & & -1 & 2 & \end{bmatrix}$	$0\le x_i\le 25$, $i=1,\ldots,d$, $d=20,50,100$	T	2
10	$f(\boldsymbol{x})=\Sigma_1^{20}x_i^2+\left[\Sigma_1^{20}\frac{1}{2}ix_i\right]^2+\left[\Sigma_1^{20}\frac{1}{2}ix_i\right]^4$	$0\le x_i\le 5$, $i=1,\ldots,20$,	T	4
11	$f(\boldsymbol{x})=\Sigma_1^{65}[y_i-[x_1\exp[-x_5t_i]+x_2\exp[-x_6(t_i-x_9)^2]+x_3\exp[-x_7(t_i-x_{10})^2]+x_4\exp[-x_9(t_i-x_{11})^2]]]^2$ $t_i=0.1(i-1),\ (i=1,\ldots,65)$, $\boldsymbol{y}=\begin{pmatrix} 1.366 & 1.191 & 1.112 & 1.013 & 0.991 & 0.885 & 0.831 & 0.847 & 0.786 & 0.725 & 0.746 \\ 0.679 & 0.608 & 0.655 & 0.616 & 0.606 & 0.602 & 0.626 & 0.651 & 0.724 & 0.649 & 0.649 \\ 0.694 & 0.644 & 0.624 & 0.661 & 0.612 & 0.558 & 0.533 & 0.495 & 0.500 & 0.423 & 0.395 \\ 0.375 & 0.372 & 0.391 & 0.396 & 0.405 & 0.428 & 0.429 & 0.523 & 0.562 & 0.607 & 0.653 \\ 0.672 & 0.708 & 0.633 & 0.668 & 0.645 & 0.632 & 0.591 & 0.559 & 0.597 & 0.625 & 0.739 \\ & 0.710 & 0.729 & 0.720 & 0.636 & 0.581 & 0.428 & 0.292 & 0.162 & 0.098 & 0.054 \end{pmatrix}$	$0\le x_1\le 1.6$; $0\le x_i\le 2$, $i=2,\ldots,5$; $2\le x_i\le 8$, $i=6,\ldots,8$; $1\le x_9\le 6$, $4.5\le x_i\le 6$, $i=10,11$;	P	6
12	$f(\boldsymbol{x})=100000\Pi_{i=1}^{11}x_i^{a_i},\ \boldsymbol{\alpha}=(-0.00133172\ \ -0.002270927\ \ -0.00248546\ \ -4.67\ \ -4.671973\ \ -0.00814\ \ -0.008092\ \ -0.005\ \ -0.000909\ \ -0.00088\ \ -0.00119)$	$0.1\le x_i\le 100$, $i=1,\ldots,11$	P	11
13	$f(\boldsymbol{x})=\Sigma_1^{14}a_i/x_i\ \ \boldsymbol{\alpha}=\begin{pmatrix} 12842.275\ \ 634.25\ \ 634.25\ \ 634.125\ \ 1268\ \ 633.875 \\ 633.75\ \ 1267\ \ 760.05\ \ 33.25\ \ 1266.25\ \ 632.875\ \ 394.46\ \ 940.838 \end{pmatrix}$	$0.0001\le x_i$, $i=1,\ldots,11$; $x_i\le 0.04$, $i=1,\ldots,5$; $x_i\le 0.03$, i $=6,\ldots,14$;	P	1

14	$f(\boldsymbol{x})=\Sigma_1^{30}\alpha_i(\boldsymbol{x})$ $\alpha_i(\boldsymbol{x})=420x_i+(i-15)^3+\Sigma_{j=1}^{30}v_{ij}[(\sin(\log(v_{ij})))^5+(\cos(\log(v_{ij})))^5]$ $v_{ij}=\sqrt{x_j^2+i/j}$	$-2\le x_i\le 2$, $i=1,\ldots,30$	T	1
15	$f(\boldsymbol{x})=\Sigma_1^{d}i(x_i^2+x_i^4)$	$-2\le x_i\le 4$, $i=1,\ldots,d$, $d=20$ *or* 50	P	1

References

[1] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., 1989, "Design and Analysis of Computer Experiments," Stat. Sci., **4**(4), pp. 409–423.

[2] Myers, R. H., and Montgomery, D., 1995, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley, Toronto, ON.

[3] Simpson, T. W., Lin, D. K. J., and Chen, W., 2001, "Sampling Strategies for Computer Experiments: Design and Analysis," Int. J. Reliab. Appl., **2**(3), pp. 209–240.

[4] Jin, R., Chen, W., and Sudjianto, A., 2005, "An Efficient Algorithm for Constructing Optimal Design of Computer Experiments," J. Stat. Plan. Infer., **134**(1), pp. 268–287.

[5] Shan, S., and Wang, G. G., 2010, "Survey of Modeling and Optimization Strategies to Solve High-Dimensional Design Problems With Computationally-Expensive Black-Box Functions," Struct. Multidiscip. Optim., **41**, pp. 219–241

[6] Chen, V. C. P., Tsui, K.-L., Barton, R. R., and Meckesheimer, M., 2006, "A Review on Design, Modeling and Applications of Computer Experiments," IIE Trans., **38**, pp. 273–291.

[7] Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., and Tucker, P. K., 2005, "Surrogate-Based Analysis and Optimization," Prog. Aerosp. Sci., **41**, pp. 1–28.

[8] Simpson, T. W., Toropov, V., and Balabanov, V., 2008, "Design and Analysis of Computer Experiments in Multidisciplinary Design Optimization: A Review of How Far We Have Come—or Not," 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, BC, Canada. Sep. 10–12, Paper No. 2208-5802.

[9] Wang, G. G., and Shan, S., 2007, "Review of Metamodeling Techniques in Support of Engineering Design Optimization," ASME J. Mech. Des., **129**, pp. 370–389.

[10] Koch, P. N., Simpson, T. W., Allen, J. K., and Mistree, F., 1999, "Statistical Approximations for Multidisciplinary Design Optimization: The Problem of Size," J. Aircr., **36**(1), pp. 275–286.

[11] Simpson, T. W., Peplinski, J., Koch, P. N., and Allen, J. K., 2001, "Metamodels for Computer-Based Engineering Design: Survey and Recommendations," Eng. Comput., **17**(2), pp. 129–150.

[12] Friedman, J. H., and Stuetzle, W., 1981, "Projection Pursuit Regression," J. Am. Stat. Assoc., **76**(376), pp. 817–823.

[13] Friedman, J. H., 1991, "Multivariate Adaptive Regressive Splines," Ann. Stat., **19**(1), pp. 1–67.

[14] Tunga, M. A., and Demiralp, M., 2006, "Hybrid High Dimensional Model Representation (HHDMR) on the Partitioned Data," J. Comput. Appl. Math., **185**(1), pp. 107–132.

[15] Rabitz, H., Alis, Ö. F., Shorter, J., and Shim, K., 1999, "Efficient Input—Output Model Representations," Comput. Phys. Commun., **117**(1–2), pp. 11–20.

[16] Rabitz, H., and Alis, Ö. F., 1999, "General Foundations of High-Dimensional Model Representations," J. Math. Chem., **25**(2–3), pp. 197–233.

[17] Shorter, J. A., Ip, P. C., and Rabitz, H. A., 1999, "An Efficient Chemical Kinetics Solver Using High Dimensional Model Representation," J. Phys. Chem. A, **103**(36), pp. 7192–7198.

[18] Li, G., Wang, S.-W., Rosenthal, C., and Rabitz, H., 2001, "High Dimensional Model Representations Generated From Low Dimensional Data Samples. I. mp-Cut-HDMR," J. Math. Chem., **30**(1), pp. 1–30.

[19] Tunga, M. A., and Demiralp, M., 2005, "A Factorized High Dimensional Model Representation on the Nodes of a Finite Hyperprismatic Regular Grid," Appl. Math. Comput., **164**(3), pp. 865–883.

[20] Shan, S., and Wang, G. G., 2009, "Metamodeling for High Dimensional Simulation-Based Design Problems," ASME J. Mech. Des., **132**, p. 051009.

[21] Booker, A. J., 1998, "Design and Analysis of Computer Experiments," Seventh AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO, pp. 118–128.

[22] Hooker, G., 2004, "Discovering Additive Structure in Black Box Functions," Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, Aug. 22–25.

[23] Wang, L., Beeson, D., Wiggs, G., and Rayasam, M., 2006, "A Comparison of Meta-Modeling Methods Using Practical Industry Requirements," 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, Newport, RI, May 1–4, AIAA Paper No. 2006-1811.

[24] Tecko, I. V., Livingstone, D. J., and Luik, A. I., 1995, "Neural Network Studies. 1. Comparison of Overfitting and Overtraining," J. Chem. Inf. Comput. Sci., **35**, pp. 826–833.

[25] Schittkowski, K., 1987, *More Test Examples for Nonlinear Programming Codes*, Springer-Verlag, New York.