J Math Chem (2012) 50:99–130
DOI 10.1007/s10910-011-9898-0

ORIGINAL PAPER

General formulation of HDMR component functions with independent and correlated variables

Genyuan Li · Herschel Rabitz

Received: 25 July 2011 / Accepted: 11 August 2011 / Published online: 25 August 2011

Abstract The High Dimensional Model Representation (HDMR) technique decomposes an n-variate function $f(\mathbf{x})$ into a finite hierarchical expansion of component functions in terms of the input variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The uniqueness of the HDMR component functions is crucial for performing global sensitivity analysis and other applications. When $x_1, x_2, \ldots, x_n$ are independent variables, the HDMR component functions are uniquely defined under a specific so called *vanishing condition*. A new formulation for the HDMR component functions is presented including cases when $\mathbf{x}$ contains correlated variables. Under a relaxed vanishing condition, a general formulation for the component functions is derived providing a unique HDMR decomposition of $f(\mathbf{x})$ for independent and/or correlated variables. The component functions with independent variables are special limiting cases of the general formulation. A novel numerical method is developed to efficiently and accurately determine the component functions. Thus, a unified framework for the HDMR decomposition of an n-variate function $f(\mathbf{x})$ with independent and/or correlated variables is established. A simple three variable model with a correlated normal distribution of the variables is used to illustrate this new treatment.

Keywords HDMR · Global sensitivity analysis · D-MORPH regression · Extended bases · Least-squares regression · Orthonormal polynomial

---

G. Li · H. Rabitz (✉)
Department of Chemistry, Princeton University, Princeton, NJ 08544, USA
e-mail: hrabitz@princeton.edu

G. Li
e-mail: genyuan@princeton.edu

1 Introduction

Many problems in science and engineering reduce to the need for efficiently constructing a map of the relationship between sets of high dimensional system input and output variables. The system may be described by a mathematical model (e.g., typically a set of differential equations), where the input variables may be initial and/or boundary conditions, parameters or functions residing in the model, and the output variables may be the solutions to the model or a functional of them. The input–output (IO) behavior may also be based on observations in the laboratory or field where a mathematical model cannot readily be constructed for the system. In this case the IO system is simply considered as a black box where the input consists of the measured laboratory or field (control) variables and the output is the observed system responses. Regardless of the circumstances, the input is often very high dimensional with many variables even if the output is only a single quantity. We refer to the input variables collectively as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ with n possibly ranging up to $\sim 10^2 - 10^3$ or more, and the output as $f(\mathbf{x})$.

High dimensional model representation (HDMR) is a general set of quantitative model assessment and analysis tools [1–15] for capturing high dimensional IO system behavior. As the impact of the multiple input variables on the output can be independent and cooperative, it is natural to express the model output $f(\mathbf{x})$ as a finite hierarchical expansion in terms of the input variables [16]:

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_i, x_j) + \cdots + f_{12\ldots n}(x_1, x_2, \ldots, x_n) \tag{1}$$

where the zeroth order component function f_0 is a constant representing the mean response to $f(\mathbf{x})$, the first order component function $f_i(x_i)$ gives the independent contribution to $f(\mathbf{x})$ by the i-th input variable acting alone, the second order component function $f_{ij}(x_i, x_j)$ gives the pair cooperative contribution to $f(\mathbf{x})$ by the input variables x_i and x_j, etc. The last term $f_{12\ldots n}(x_1, x_2, \ldots, x_n)$ contains any residual n-th order cooperative contribution of all the input variables. Thus, the above HDMR expansion with a finite number of terms is always exact.

The basic conjecture underlying HDMR is that the component functions in Eq. 1 arising in typical real problems are likely to reflect only low order cooperativity among the input variables. In particular, experience shows that an HDMR expansion to 2nd order

$$f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_i, x_j) \tag{2}$$

often provides a satisfactory description of $f(\mathbf{x})$ for many high dimensional systems when the input variables are properly chosen and the component functions are optimally constructed.

Exploiting the expected low order variable cooperativity in high dimensional systems can only be done if practical formulations for constructing the HDMR component functions can be found. In this regard, two classes of problems arise: either all of the input variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are independent or at least some portion of the variables in $\mathbf{x}$ are correlated. Standard formulations of HDMR deal with the case of independent variables, and this paper makes proper extensions to treat correlated variables.

As background, when $x_1, x_2, \ldots, x_n$ are independent variables (i.e., statistically uncorrelated), the component functions f_0, $f_i(x_i)$, $f_{ij}(x_i, x_j), \ldots$ can be *optimally and uniquely* defined for a particular $f(\mathbf{x})$ over the entire domain of $\mathbf{x}$ ($[a_i, b_i]$, $i = 1, 2, \ldots, n$) by imposing the *vanishing condition* [10]

$$\int_{a_i}^{b_i} w_s(x_s) f_{i_1 i_2 \ldots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \mathrm{d}x_s = 0, \qquad \forall s \in \{i_1, i_2, \ldots, i_k\} \tag{3}$$

where

$$\begin{cases} w_i(x_i) \geq 0, & (a_i \leq x_i \leq b_i), \\ \int_{a_i}^{b_i} w_i(x_i) \mathrm{d}x_i = 1, & (i = 1, 2, \ldots, n). \end{cases} \tag{4}$$

The resultant component functions have the form

$$f_0 = \int w(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{5}$$

$$f_i(x_i) = \int w_{-i}(\mathbf{x}_{-i}) f(x_i, \mathbf{x}_{-i}) \mathrm{d}\mathbf{x}_{-i} - f_0, \tag{6}$$

$$\begin{aligned} f_{ij}(x_i, x_j) &= \int w_{-ij}(\mathbf{x}_{-ij}) f(x_i, x_j, \mathbf{x}_{-ij}) \mathrm{d}\mathbf{x}_{-ij} \\ &- f_i(x_i) - f_j(x_j) - f_0, \\ &\ldots \end{aligned} \tag{7}$$

Here $w(\mathbf{x})$ is the probability density function (pdf) for $\mathbf{x}$ satisfying

$$w(\mathbf{x}) = \prod_{i=1}^{n} w_i(x_i), \tag{8}$$

and $\mathbf{x}_{-i}$ and $\mathbf{x}_{-ij}$ refer to $\mathbf{x}$ without the elements x_i and x_i, x_j, respectively. Similarly, $w_{-i}(\mathbf{x}_{-i})$ and $w_{-ij}(\mathbf{x}_{-ij})$ are marginal pdf's obtained by integrating the distribution over x_i and x_i, x_j, respectively:

$$w_{-i}(\mathbf{x}_{-i}) = \int w(\mathbf{x})\mathrm{d}x_i = \prod_{\substack{k=1\\k\neq i}}^{n} w_k(x_k), \tag{9}$$

$$w_{-ij}(\mathbf{x}_{-ij}) = \int w(\mathbf{x})\mathrm{d}x_i\mathrm{d}x_j = \prod_{\substack{k=1\\k\neq i,j}}^{n} w_k(x_k). \tag{10}$$

For the sake of notational neatness, we omit the specific integration dimension and range and use $\int$ to represent all integrations. As the expression for a component function of a particular order only depends on the lower order component functions (e.g., see Eq. 7), all the HDMR component functions can be determined sequentially starting from f_0. The last term $f_{12\ldots n}(x_1, x_2, \ldots, x_n)$ is given from the difference between $f(\mathbf{x})$ and the sum of all the other component functions. The condition in Eq. 3 assures that the HDMR component functions are mutually orthogonal

$$\begin{gathered}\int w(\mathbf{x}) f_{i_1i_2\cdots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) f_{j_1j_2\cdots j_l}(x_{j_1}, x_{j_2}, \ldots, x_{j_l})\mathrm{d}\mathbf{x} = 0.\\ \{i_1, i_2, \cdots, i_k\} \neq \{j_1, j_2, \ldots, j_l\}\end{gathered} \tag{11}$$

For a uniform input distribution (i.e., $w_i(x_i) = 1$) and $[a_i, b_i] = [0, 1]$, Eqs. 5–7 reduce to the formulas given by Sobol [16].

For independent variables (i.e., Eq. 8), the conditional pdf is equal to its corresponding marginal pdf. For example, the conditional pdf for a given value of x_i is

$$\begin{aligned} w_{\mathbf{x}_{-i}|x_i}(\mathbf{x}_{-i}) &= w(x_i, \mathbf{x}_{-i})/w_i(x_i) = w(\mathbf{x})/w_i(x_i)\\ &= \prod_{\substack{k=1\\k\neq i}}^{n} w_k(x_k) = w_{-i}(\mathbf{x}_{-i}), \end{aligned} \tag{12}$$

and similarly for a given value of x_i, x_j

$$\begin{aligned} w_{\mathbf{x}_{-ij}|x_i,x_j}(\mathbf{x}_{-ij}) &= w(x_i, x_j, \mathbf{x}_{-ij})/w_{ij}(x_i, x_j) = w(\mathbf{x})/(w_i(x_i)w_j(x_j))\\ &= \prod_{\substack{k=1\\k\neq i,j}}^{n} w_k(x_k) = w_{-ij}(\mathbf{x}_{-ij}). \end{aligned} \tag{13}$$

In this case, Eqs. 5–7 can then be expressed in another form

$$f_0 = E(f(\mathbf{x})), \tag{14}$$

$$f_i(x_i) = E(f(\mathbf{x})|x_i) - f_0, \tag{15}$$

$$f_{ij}(x_i, x_j) = E(f(\mathbf{x})|x_i, x_j) - f_i(x_i) - f_j(x_j) - f_0, \tag{16}$$

$$\ldots$$

where E denotes the expected value. Conditional expectation can be formulated as a *regression* process [17]. Therefore, the component functions in Eqs. 14–16 may be determined by any parametric or non-parametric fitting technique (e.g., least-squares regression) with suitable experimental or modeling data. When the variables $x_1, x_2, \ldots, x_n$ are independent and putting aside the influence of data errors, then either individual or simultaneous determination of the component functions, $f_i(x_i)$, $f_{ij}(x_i, x_j)$, … by minimizing the squared error will give the same results. Equations 14–16 are commonly used under these circumstances.

Distinct, but formally equivalent HDMR expansions for independent variables, all of the same structure as Eq. 1, have been constructed. There are two commonly used HDMR expansions: Cut- and RS (Random Sampling)-HDMR. Cut-HDMR expresses $f(\mathbf{x})$ in reference to a specified cut point $\bar{\mathbf{x}}$ in the input domain, and is useful for data sampling performed in a controllable orderly fashion [1–3,5,11]. RS-HDMR depends on the integration of $f(\mathbf{x})$ over the whole input domain, and applies to data sampled under an arbitrary pdf, as is often the case with laboratory and field data [6–10,12–15]. Various practical approaches and modifications of HDMR, e.g., mp-Cut-HDMR [5], Multicut-HDMR [11], lp-RS-HDMR [9], etc., have been developed to improve the accuracy and efficiency of HDMR.

Constructing the HDMR component functions only requires the values of the output $f(\mathbf{x})$ at a suitable set of points $\mathbf{x}$. These values may be obtained either from observational data (without knowing the internal structure of the system) or from a computational model. Once the component functions are determined, their evaluation only involves simple algebraic calculations, and calling up $f(\mathbf{x})$ with HDMR for a new (previously unconsidered) point $\mathbf{x}$ is extremely fast. Therefore, the resultant HDMR expansion can be used as a Fully Equivalent Operational Model (FEOM) generated either from measured laboratory/field data or a mathematical model. A reliable FEOM can replace time-consuming further modeling or experiments to significantly reduce the computational or laboratory/field measurement effort (e.g., to facilitate optimization and inversion problems where model calculations and data collection are called repeatedly). Moreover, RS-HDMR provides a means to enable the performance of global sensitivity analysis to systematically assess which inputs, over a global domain, have the most influence along with their patterns of independent and cooperative actions upon the output [15]. In applications, it is important to distinguish (i) cooperative behavior of the variables in $\mathbf{x}$ caused by their role in the physical map $\mathbf{x} \rightarrow f$ (e.g., through a differential equation) verses that of (ii) inherent statistical correlation arising in the pdf of the variables $\mathbf{x}$ themselves.

For correlated variables $\mathbf{x}$, the pdf $w(\mathbf{x})$ is not separable as a product $\prod_{i=1}^{n} w_i(x_i)$. The conditional expectation of $f(\mathbf{x})$ for a *given set* of variables can also depend on *other* correlated variables (i.e., there can be a contribution from other correlated variables). Thus, with correlated variables $f(\mathbf{x})$ is no longer the sum of the component functions given by Eqs. 14–16 [18]. In addition, least-squares regression for determining the component functions individually or simultaneously may give different results. As $w(\mathbf{x})$ is not separable for correlated variables, the vanishing condition, Eq. 3, cannot be used as a basis to determine the HDMR component functions. Without a proper vanishing condition the HDMR decomposition may not be unique for $f(\mathbf{x})$ with correlated variables. Importantly, a global sensitivity analysis based on these non-unique component functions is not meaningful.

Friedman [19] noted that taking the marginal distribution rather than a conditional distribution given a subset of $\mathbf{x}$ preserves the additive structure of $f(\mathbf{x})$, which is the basis for the unique decomposition of an n-variate function $f(\mathbf{x})$ with correlated variables. Hooker [20] relaxed the vanishing condition, Eq. 3, and then proved that the decomposition in Eq. 1 for an n-variate function $f(\mathbf{x})$ obtained by minimization of the squared error will be unique.

Using the relaxed vanishing condition given by Hooker, we will derive general formulas for the component functions to produce a unique HDMR decomposition of $f(\mathbf{x})$ with independent and/or correlated variables. We will prove that the formulas currently used for independent variables are a special case of the general formulas. We will also develop a novel numerical method to efficiently and accurately determine the unique component functions. Thus, a unified framework for a unique HDMR decomposition of an n-variate function $f(\mathbf{x})$ with independent and/or correlated variables will be established.

The paper is organized as follows. Section 2 derives the general formulas specifying the unique HDMR component functions. Section 3 presents the numerical method for determining the component functions based on extended bases [13] and D-MORPH (Diffeomorphic Modulation under Observable Response Preserving Homotopy) regression [21]. In Sect. 4, a three variable model with correlation amongst the variables is used to illustrate the new formulation of HDMR as well as the performance of the numerical method. Finally, Sect. 5 contains concluding remarks.

2 General formulas for HDMR component functions

In order to make the formulation tractable we use the following multi-index notation [20]. Given the subset $u \subseteq \{1, 2, \ldots, n\}$, we denote by $\mathbf{x}_u$ those variables whose indexes are in u, and $\mathbf{x}_{-u}$ denotes the subset of variables with indexes not in u. We will also write $u \subseteq n$ in place of $u \subseteq \{1, 2, \ldots, n\}$ for simplicity. Note that u includes the empty set $\emptyset$. Then, the HDMR expansion, Eq. 1, can be written in brief form as

$$f(\mathbf{x}) = \sum_{u \subseteq n} f_u(\mathbf{x}_u). \tag{17}$$

Since $\emptyset \in u$, the above summation contains f_0.

Hooker defines the HDMR component functions $\{f_u(\mathbf{x}_u | u \subseteq n)\}$ for $f(\mathbf{x})$ with correlated variables as satisfying

$$\{f_u(\mathbf{x}_u | u \subseteq n)\} = \operatorname{argmin}_{\{g_u \in \mathbb{L}^2(\mathbb{R}^u), u \subseteq n\}} \int \left(\sum_{u \subseteq n} g_u(\mathbf{x}_u) - f(\mathbf{x}) \right)^2 w(\mathbf{x}) d\mathbf{x} \tag{18}$$

under a relaxed vanishing condition

$$\forall u \subseteq n, \forall i \in u, \int f_u(\mathbf{x}_u) w(\mathbf{x}) \mathrm{d}x_i \mathrm{d}\mathbf{x}_{-u} = 0. \tag{19}$$

This criterion is equivalent to the *hierarchical orthogonality* condition

$$\forall v \subset u, \forall g_v : \int f_u(\mathbf{x}_u) g_v(\mathbf{x}_v) w(\mathbf{x}) \mathrm{d}\mathbf{x} = \langle f_u(\mathbf{x}_u), g_v(\mathbf{x}_v) \rangle = 0, \tag{20}$$

i.e., *a component function is only required to be orthogonal to all nested lower order component functions whose variables are a subset of its variables.* For example, $f_{ijk}(x_i, x_j, x_k)$ is only required to be orthogonal to $f_i(x_i)$, $f_j(x_j)$, $f_k(x_k)$, $f_{ij}(x_i, x_j)$, $f_{ik}(x_i, x_k)$, and $f_{jk}(x_j, x_k)$.

Hooker proved that when the support of $w(\mathbf{x})$ is grid closed, for any $\{g_u | u \subseteq n\} \in \mathbb{L}^2$ with at least one $g_u \neq 0$ that satisfy Eq. 19, the set of functions $\{g_u | u \subseteq n\}$ are linearly independent under the inner product defined by $w(\mathbf{x})$. Then Eq. 18 has a unique minimizer $f_u(\mathbf{x}_u | u \subseteq n)$ under the condition Eq. 19; otherwise, $\{g_u | u \subseteq n\}$ would be linearly dependent.

The name "grid closed" means the existence of a grid for every point $\mathbf{x}$ in the desired domain Ω,[1] which is a reasonable circumstance in most realistic cases. Therefore, *if the* HDMR *component functions can be constructed by minimization of the squared error under the vanishing condition,* Eq. 19, *or the equivalent hierarchical orthogonal condition,* Eq. 20 (Eq. 19 is the necessary and sufficient condition for Eq. 20), *the resultant component functions* $\{f_u(\mathbf{x}_u | u \subseteq n)\}$ *are unique.*

Equation 19 can be simplified by implicitly performing the integration for $\mathbf{x}_{-u}$, which gives

$$\forall u \subseteq n, \forall i \in u, \int f_u(\mathbf{x}_u) w_u(\mathbf{x}_u) \mathrm{d}x_i = 0, \tag{21}$$

where $w_u(\mathbf{x}_u)$ is the marginal pdf for $\mathbf{x}_u$. Equation 21 is more convenient than Eq. 19 in the following treatment. It can be readily proved that for independent variables, i.e., $w(\mathbf{x}) = \prod_{i=1}^{n} w_i(x_i)$, Eq. 21 reduces to Eq. 3. This result implies that Eq. 19 (or Eq. 21) is a general vanishing condition which includes Eq. 3 as a special case for independent variables.

We may use the vanishing condition, Eq. 19 or the equivalent Eq. 21, to uniquely specify the HDMR component functions. The operations involve multiplying Eq. 1 (or equivalently Eq. 17) by an appropriate marginal pdf, integration over the associated subset of variables, and application of the vanishing condition in Eq. 21. The procedure will be specifically presented for f_0, $f_i(x_i)$, $f_{ij}(x_i, x_j)$ and $f_{ijk}(x_i, x_j, x_k)$. For f_0, $w(\mathbf{x})$ is used and f_0 is the same as that for independent variables

$$f_0 = \int f(\mathbf{x}) w(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{22}$$

[1] Consider a simple example: $f(x_1, x_2)$. The domain Ω for (x_1, x_2) is said to be grid closed if for a point (x_1, x_2) there always exists another point with only one of the variables being the same. For instance, the line $x_1 = x_2$ is not grid closed; in this restrictive case there does not exist another point with only one variable being the same.

For $f_i(x_i)$, the marginal pdf $w_{-i}(\mathbf{x}_{-i})$ is used, and

$$\begin{aligned}\int f(\mathbf{x})w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i} &= \sum_{u\subseteq n}\int f_u(\mathbf{x}_u)w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i}\\ &= f_0\int w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i} + f_i(x_i)\int w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i}\\ &\quad + \sum_{\{i\}\subset u\subseteq n}\int f_u(\mathbf{x}_u)w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i}\\ &\quad + \sum_{i\notin u\subseteq n}\int f_u(\mathbf{x}_u)w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i}\\ &= f_0 + f_i(x_i) + \sum_{\{i\}\subset u\subseteq n}\int f_u(\mathbf{x}_u)w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i}. \end{aligned} \tag{23}$$

Here $\{i\}$ denotes the subset only containing element i. The relations

$$\int w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i} = 1 \tag{24}$$

and

$$\sum_{i\notin u\subseteq n}\int f_u(\mathbf{x}_u)w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i} = \sum_{i\notin u\subseteq n}\int f_u(\mathbf{x}_u)w_u(\mathbf{x}_u)\mathrm{d}\mathbf{x}_u = 0 \tag{25}$$

were used in the above operations. Equation 25 is zero because each integral satisfies Eq. 21. Thus we have

$$\begin{aligned} f_i(x_i) &= \int f(\mathbf{x})w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i} - f_0 - \sum_{\{i\}\subset u\subseteq n}\int f_u(\mathbf{x}_u)w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i}\\ &= \int f(\mathbf{x})w_{-i}(\mathbf{x}_{-i})\mathrm{d}\mathbf{x}_{-i} - f_0 - h_i(x_i). \end{aligned} \tag{26}$$

The last term $h_i(x_i)$ depends on all of the component functions, higher than first order, containing x_i.

The procedure is similar for $f_{ij}(x_i, x_j)$ and $f_{ijk}(x_i, x_j, x_k)$, but with the marginal pdf's $w_{-ij}(\mathbf{x}_{-ij})$ and $w_{-ijk}(\mathbf{x}_{-ijk})$, respectively. We find

$$\begin{aligned} f_{ij}(x_i, x_j) &= \int f(\mathbf{x})w_{-ij}(\mathbf{x}_{-ij})\mathrm{d}\mathbf{x}_{-ij} - f_0 - f_i(x_i) - f_j(x_j)\\ &\quad - \sum_{\substack{u\subseteq n\\ \{i,j\}\bigcap u\neq\emptyset}}\int f_u(\mathbf{x}_u)w_{-ij}(\mathbf{x}_{-ij})\mathrm{d}\mathbf{x}_{-ij}\\ &= \int f(\mathbf{x})w_{-ij}(\mathbf{x}_{-ij})\mathrm{d}\mathbf{x}_{-ij} - f_0 - f_i(x_i) - f_j(x_j) - h_{ij}(x_i, x_j). \end{aligned} \tag{27}$$

The last term $h_{ij}(x_i, x_j)$ depends on all the component functions of 2nd and higher order containing either x_i, or x_j, or both x_i, x_j. For the 3rd order component function we have

$$\begin{aligned} f_{ijk}(x_i, x_j, x_k) &= \int f(\mathbf{x}) w_{-ijk}(\mathbf{x}_{-ijk}) \mathrm{d}\mathbf{x}_{-ijk} - f_0 - f_i(x_i) - f_j(x_j) \\ &\quad - f_{ij}(x_i, x_j) - f_{ik}(x_i, x_k) - f_{jk}(x_j, x_k) \\ &\quad - \sum_{\substack{u \subseteq n \\ \{i,j,k\} \bigcap u \neq \emptyset}} \int f_u(\mathbf{x}_u) w_{-ijk}(\mathbf{x}_{-ijk}) \mathrm{d}\mathbf{x}_{-ijk} \\ &= \int f(\mathbf{x}) w_{-ijk}(\mathbf{x}_{-ijk}) \mathrm{d}\mathbf{x}_{-ijk} - f_0 - f_i(x_i) - f_j(x_j) \\ &\quad - f_{ij}(x_i, x_j) - f_{ik}(x_i, x_k) - f_{jk}(x_j, x_k) - h_{ijk}(x_i, x_j, x_k). \end{aligned} \tag{28}$$

The last term $h_{ijk}(x_i, x_j, x_k)$ in Eq. 28 depends on the component functions of 3rd and higher order containing any or all of the variables x_i, x_j, x_k.

Equations 26–28 show that for correlated variables, a particular component function of order l depends on *all* other component functions that contain one or more of the same variables (i.e., respectively in $h_i(x_i)$, $h_{ij}(x_i, x_j)$, $h_{ijk}(x_i, x_j, x_k)$). Thus, it is impossible to determine the component functions sequentially starting from f_0. For independent variables, the last h-terms in Eqs. 26–28 vanishes because $w_{-ij}(\mathbf{x}_{-ij})$, $w_{-ijk}(\mathbf{x}_{-ijk})$ are products of $w_l(x_l)(l \in \mathbf{x}_{-ij}, \mathbf{x}_{-ijk})$, and at least for one variable the vanishing condition, Eq. 3, is satisfied. Then the formulas reduce to Eqs. 5–7. Therefore, Eqs. 26–28 are general for independent and/or correlated variables.

Despite the coupled nature of the component functions with correlated variables, approximate formulas may be constructed. When $f(\mathbf{x})$ can be exactly represented by a low order HDMR expansion, even exact formulas may be obtained in some cases. An illustration will be given in Sect. 4.

3 Practical determination of component functions: extended bases and D-MORPH regression

The mathematical form of $f(\mathbf{x})$ is often unknown in realistic applications. Even if a formula for $f(\mathbf{x})$ is available, the integrations in Eqs. 26–28 may not be possible analytically. Thus, direct utilization of Eqs. 26–28 rarely can be employed for constructing the component functions. In typical applications, the function $f(\mathbf{x})$ is only available by sampling points in $\mathbf{x}$ either from modeling or experiments. Therefore, a practical numerical method is needed to construct each unique component function *directly from minimizing the squared error under the relaxed vanishing condition* (Eqs. 18, 19) *or minimizing the squared error under the hierarchical orthogonality condition* (Eqs. 18, 20). Hooker suggested a way to determine these component functions with $\mathbf{x}$ sampled on a grid [20]. In this fashion, Eqs. 18 and 19 become a high dimensional, sparse linear algebraic system of equations. This procedure appears to be computational demanding. We take another route: minimization of the squared error under the hierarchical orthogonality condition, which can be realized easily. We propose an efficient method to determine the unique component functions based on utilizing two tools: extended bases [13] and D-MORPH regression [21], as explained below.

3.1 Extended bases

In previous work [6,7,10,13], the HDMR component functions were approximated by expansions in particular basis functions. The *sufficient* condition for *hierarchical orthogonality* of the component functions is that the subspace of the Hilbert space spanned by the basis functions for any lower order component function is a normal subspace of the subspace spanned by the basis functions of the nested higher order component functions. Suppose that a subspace V in Hilbert space is spanned by the basis $\{v_1, v_2, \ldots, v_k\}$, and a larger subspace $U (\supset V)$ is spanned by the extended basis $\{v_1, v_2, \ldots, v_k, v_{k+1}, \ldots, v_m\}$. Then U can be decomposed as

$$U = V \oplus V^{\perp}$$

where $V^{\perp}$ is the orthogonal complement subspace of V in U. One can always find a vector in $V^{\perp}$ (i.e., a certain linear combination of $v_1, v_2, \ldots, v_k, v_{k+1}, \ldots, v_m$) orthogonal to all vectors in V [22].

To satisfy this sufficient condition, the component functions are approximated by expansions in some suitable basis functions $\{\varphi\}$ (polynomials, splines, etc.) as follows [13]

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^{(0)i} \varphi_r^i(x_i), \tag{29}$$

$$\begin{aligned} f_{ij}(x_i, x_j) \approx & \sum_{r=1}^{k} \left[\alpha_r^{(ij)i} \varphi_r^i(x_i) + \alpha_r^{(ij)j} \varphi_r^j(x_j)\right] \\ & + \sum_{p=1}^{l} \sum_{q=1}^{l} \beta_{pq}^{(0)ij} \varphi_p^i(x_i) \varphi_q^j(x_j), \end{aligned} \tag{30}$$

$$\begin{aligned} f_{ijk}(x_i, x_j, x_k) \approx & \sum_{r=1}^{k} \left[\alpha_r^{(ijk)i} \varphi_r^i(x_i) + \alpha_r^{(ijk)j} \varphi_r^j(x_j) + \alpha_r^{(ijk)k} \varphi_r^k(x_k)\right] \\ & + \sum_{p=1}^{l} \sum_{q=1}^{l} \left[\beta_{pq}^{(ijk)ij} \varphi_p^i(x_i) \varphi_q^j(x_j) + \beta_{pq}^{(ijk)ik} \varphi_p^i(x_i) \varphi_q^k(x_k) \right. \\ & \left. + \beta_{pq}^{(ijk)jk} \varphi_p^j(x_j) \varphi_q^k(x_k)\right] \end{aligned}$$

$$+\sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\gamma_{pqr}^{(0)ijk}\varphi_p^i(x_i)\varphi_q^j(x_j)\varphi_r^k(x_k), \tag{31}$$
$$\cdots$$

where k, l, m are integers. Note that in Eqs. 29–31 the basis functions of the lower order component functions are always a subset of those for the higher order ones, and the hierarchical orthogonality between $f_i(x_i) f_{ij}(x_i, x_j),\ f_{ijk}(x_i, x_j, x_k), \ldots$ can be readily achieved. For example, consider $f_{ij}(x_i, x_j)$. One can always find suitable values for $\alpha_r^{(ij)i}$, $\alpha_r^{(ij)j}$ and $\beta_{pq}^{(0)ij}$ such that $f_{ij}(x_i, x_j)$ is orthogonal to $\varphi_r^i(x_i)$, $\varphi_r^j(x_j)$, as well as any linear combinations of them, and consequently, $f_i(x_i)$ and $f_j(x_j)$ as demanded by Eq. 20.

The optimal orthonormal polynomial basis $\{\varphi\}$ satisfies [13]

$$\int w_i(x_i)\varphi_r^i(x_i)\mathrm{d}x_i \approx \sum_{s=1}^{N}\varphi_r^i(x_i^{(s)})/N = 0, \text{ for all } \quad r, i \tag{32}$$

$$\int w_i(x_i)(\varphi_r^i(x_i))^2\mathrm{d}x_i \approx \sum_{s=1}^{N}(\varphi_r^i(x_i^{(s)}))^2/N = 1, \quad \text{for all} \quad r, i \tag{33}$$

$$\int w_i(x_i)\varphi_p^i(x_i)\varphi_q^i(x_i)\mathrm{d}x_i \approx \sum_{s=1}^{N}\varphi_p^i(x_i^{(s)})\varphi_q^i(x_i^{(s)})/N = 0. \quad p \neq q \tag{34}$$

In this fashion the basis may be constructed from a set of data generated according to a given pdf, where $x_i^{(s)}$ is the s-th sample and N is the total number of samples. The basis set members have zero mean, unit norm and are mutually orthogonal with respect to the marginal pdf weight $w_i(x_i)$. In many cases, satisfactory accuracy is likely attainable using only $\varphi_r^i(x_i), r \leq 3$ to approximate $f_i(x_i),\ f_{ij}(x_i, x_j)$ and $f_{ijk}(x_i, x_j, x_k)$.

Employing the formulas in Eqs. 29–31, the 3rd order HDMR expansion for an n-variate function $f(\mathbf{x})$ can be expressed as

$$\begin{aligned} f(\mathbf{x}) \approx f_0 &+ \sum_{i=1}^{n}\sum_{r=1}^{k}\Big(\alpha_r^{(0)i} + \sum_{\substack{j=1\\ j\neq i}}^{n}\alpha_r^{(ij)i} + \sum_{\substack{j<k=1\\ j,k\neq i}}^{n}\alpha_r^{(ijk)i}\Big)\varphi_r^i(x_i) \\ &+ \sum_{1\leq i<j\leq n}\sum_{p=1}^{l}\sum_{q=1}^{l}\left(\beta_{pq}^{(0)ij} + \sum_{k=1\, k\neq i,j}^{n}\beta_{pq}^{(ijk)ij}\right)\varphi_p^i(x_i)\varphi_q^j(x_j) \\ &+ \sum_{1\leq i<j<k\leq n}\sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\gamma_{pqr}^{(0)ijk}\varphi_p^i(x_i)\varphi_q^j(x_j)\varphi_r^k(x_k). \end{aligned} \tag{35}$$

The constant coefficients $\{\alpha\}$, $\{\beta\}$, $\{\gamma\}$ may be determined by minimizing the squared error (e.g., least-squares regression) from the data $(\mathbf{x}^{(s)}, f(\mathbf{x}^{(s)}), s = 1, 2, \ldots, N)$ generated according to the pdf $w(\mathbf{x})$.

Equation 35 can be written in vector form for all of the data

$$\phi(\mathbf{x}^{(s)})^T \mathbf{c} = f(\mathbf{x}^{(s)}) - f_0, \qquad (s = 1, 2, \ldots, N) \tag{36}$$

where

$$f_0 = \sum_{s=1}^{N} f(\mathbf{x}^{(s)})/N. \tag{37}$$

Here $\mathbf{c}$ is composed of all the unknown constant coefficients

$$\mathbf{c}^T = (\alpha_1^{(0)1} \alpha_2^{(0)1} \cdots \alpha_k^{(0)1} \alpha_1^{(0)2} \cdots \alpha_1^{(12)1} \cdots \gamma_{mmm}^{(0)(n-2)(n-1)n}), \tag{38}$$

and $\phi(\mathbf{x}^{(s)})^T$ consists of the corresponding basis functions

$$\begin{aligned}\phi(\mathbf{x}^{(s)})^T &= (\varphi_1^1(x_1^{(s)})\ \varphi_2^1(x_1^{(s)})\ \cdots\ \varphi_k^1(x_1^{(s)})\ \varphi_1^2(x_2^{(s)})\ \cdots\ \varphi_1^1(x_1^{(s)}) \\ &\quad \cdots\ \varphi_m^{(n-2)}(x_{n-2}^{(s)})\varphi_m^{(n-1)}(x_{n-1}^{(s)})\varphi_m^n(x_n^{(s)})) \\ &= (r_1(\mathbf{x}^{(s)})\ r_2(\mathbf{x}^{(s)})\ \cdots\ r_t(\mathbf{x}^{(s)})).\end{aligned} \tag{39}$$

To simplify the notation, the symbol $r_j(\mathbf{x})$ is used to represent all the above basis functions, and t is the total number of unknown coefficients. Note that there are many repetitive appearances of some basis functions in $\phi(\mathbf{x}^{(s)})^T$, because the same basis function is associated with different coefficients (see Eq. 35), i.e., some of the r_j's are the same.

Equation 36 can be written in matrix form as

$$\mathbf{\Phi c} = \mathbf{b} \tag{40}$$

where $\mathbf{\Phi}$ is an $N \times t$ matrix (the s-th row of $\mathbf{\Phi}$ is $\phi(\mathbf{x}^{(s)})^T$), and $\mathbf{b}$ is an N-dimensional vector whose s-th element is $f(\mathbf{x}^{(s)}) - f_0$. Since $\phi(\mathbf{x}^{(s)})^T$ has repeated elements, some columns of $\mathbf{\Phi}$ are identical.

The vector $\mathbf{c}$ minimizing the squared error is the solution of the normal equation for least-squares regression of Eq. 40:

$$\mathbf{\Phi}^T \mathbf{\Phi c} = \mathbf{\Phi}^T \mathbf{b} \tag{41}$$

with $\mathbf{\Phi}^T \mathbf{\Phi}$ being a $t \times t$ matrix. Dividing both sides of Eq. 41 by N will not change the solution for $\mathbf{c}$,

$$\frac{1}{N}\mathbf{\Phi}^T \mathbf{\Phi c} = \frac{1}{N}\mathbf{\Phi}^T \mathbf{b}, \tag{42}$$

and now the (i, j)th entry of $\mathbf{\Phi}^T \mathbf{\Phi}/N$ can be viewed as an approximation to the inner product of $r_i(\mathbf{x})$ and $r_j(\mathbf{x})$

$$(\mathbf{\Phi}^T \mathbf{\Phi})_{ij}/N = \sum_{s=1}^{N} r_i(\mathbf{x}^{(s)}) r_j(\mathbf{x}^{(s)})/N \approx \langle r_i(\mathbf{x}), r_j(\mathbf{x})\rangle. \tag{43}$$

As $\mathbf{\Phi}^T$ has duplicate rows, some equations in Eq. 42 are identical. These duplicate equations are redundant and can be removed. The resultant linear algebraic equation system is

$$A\mathbf{c} = \mathbf{d} \tag{44}$$

where A and $\mathbf{d}$ are just $\mathbf{\Phi}^T \mathbf{\Phi}/N$ and $\mathbf{\Phi}^T \mathbf{b}/N$ after removing the duplicate rows. Now A is a $p \times t (p < t)$ rectangular matrix. In Eq. 44 the number (t) of unknown coefficients is larger than the number (p) of equations. Such a system is consistent and has an infinite number of solutions for $\mathbf{c}$ with the general form

$$\mathbf{c} = A^+\mathbf{d} + (\mathrm{I}_t - A^+ A)\mathbf{u}, \tag{45}$$

where I_t is the identity matrix with dimension t and $\mathbf{u}$ is an arbitrary vector in $\Re^t$, and A^+ is the generalized inverse G of A satisfying all four Penrose conditions [23]

$$\begin{array}{ll} (1)\ AGA = A, & (2)\ GAG = G, \\ (3)\ (AG)^T = AG, & (4)\ (GA)^T = GA. \end{array} \tag{46}$$

The solution with the smallest norm $\|\mathbf{c}\|$ commonly produced by least-squares regression is unique and given by

$$\mathbf{c} = A^+\mathbf{d}. \tag{47}$$

This solution will be used later for comparison with the unique HDMR component functions.

The infinite number of solutions for $\mathbf{c}$ given in Eq. 45 provides the possibility to search for a solution of $\mathbf{c}$ *both minimizing the squared error and satisfying the hierarchical orthogonality condition*.

3.2 D-MORPH regression

D-MORPH (Diffeomorphic Modulation under Observable Response Preserving Homotopy) is a model exploration method, originally developed for optimal control with differential equation models [24–27]. The method was extended to the regression treatment of a model described as a linear superposition of basis functions with unknown parameters being the expansion coefficients [21]. When there are more unknown parameters than equations, the corresponding linear algebraic equation system is consistent, and has an infinite number of solutions. D-MORPH regression is a practical systematic means to search for a solution satisfying extra requirements within the infinite number of possible solutions. Equation 44 is such a system and the hierarchical orthogonality of the component functions is the extra requirement.

3.2.1 Principles of D-MORPH regression

The key principles of D-MORPH regression are briefly summarized here; further details can be found in [21]. All the solutions $\mathbf{c}$ of Eq. 44 given by Eq. 45 compose a completely connected submanifold $\mathcal{M} \subset \Re^t$. D-MORPH regression searches for a solution satisfying an extra requirement by considering an exploration path $\mathbf{c}(s)$ within $\mathcal{M}$ with s in $[0, \infty)$, which satisfies a differential equation[2]

$$\frac{\mathrm{d}\mathbf{c}(s)}{\mathrm{d}s} = P\mathbf{v}(s) = (\mathrm{I}_t - A^+ A)\mathbf{v}(s), \tag{48}$$

where P is an orthogonal projector [23] satisfying

$$P^2 = P, \qquad P^T = P, \tag{49}$$

which yields

$$P = P^2 = P^T P. \tag{50}$$

The function vector $\mathbf{v}(s)$ may be freely chosen to not only enable broad choices for exploring $\mathbf{c}(s)$, but to also continuously reduce a defined cost $\mathcal{K}(\mathbf{c}(s))$ (e.g., the model variance, fitting smoothness, the weighted norm of $\mathbf{c}$, or particularly here the hierarchical orthogonality of the component functions) along the exploration path. If the free function vector is chosen as

$$\mathbf{v}(s) = -\frac{\partial \mathcal{K}(\mathbf{c}(s))}{\partial \mathbf{c}}, \tag{51}$$

then we obtain

$$\begin{aligned}\frac{\mathrm{d}\mathcal{K}(\mathbf{c}(s))}{\mathrm{d}s} &= \left(\frac{\partial \mathcal{K}(\mathbf{c}(s))}{\partial \mathbf{c}}\right)^T \frac{\mathrm{d}\mathbf{c}(s)}{\mathrm{d}s} = \left(\frac{\partial \mathcal{K}(\mathbf{c}(s))}{\partial \mathbf{c}}\right)^T P\mathbf{v}(s) \\ &= -\left(P\frac{\partial \mathcal{K}(\mathbf{c}(s))}{\partial \mathbf{c}}\right)^T \left(P\frac{\partial \mathcal{K}(\mathbf{c}(s))}{\partial \mathbf{c}}\right) \le 0,\end{aligned} \tag{52}$$

i.e., the cost $\mathcal{K}$, used as an additional requirement, will be continuously reduced (systematically refining the model) over the course of traversing $s \ge 0$. Therefore,

$$\mathbf{c}_\infty = \lim_{s\to\infty} \mathbf{c}(s)$$

[2] $\mathbf{v}(s)$ in Eq. 48 is related with $\mathbf{u}$ in Eq. 45 through $\mathbf{v}(s) = \mathrm{d}\mathbf{u}/\mathrm{d}s$.

is the solution which minimizes $\mathcal{K}$. When the cost function is defined as a quadratic form in $\mathbf{c}$

$$\mathcal{K} = \frac{1}{2}\mathbf{c}^T B\mathbf{c}, \tag{53}$$

where B is symmetric and non-negative definite, the analytical form of $\mathbf{c}_\infty$ can be obtained as

$$\mathbf{c}_\infty = V_{t-r}(U_{t-r}^T V_{t-r})^{-1} U_{t-r}^T A^+\mathbf{d}. \tag{54}$$

where U_{t-r}, and V_{t-r} are the last $t-r$ columns of U and V obtained by singular value decomposition of PB [28]

$$PB = U\begin{bmatrix} S_r & 0 \\ 0 & 0 \end{bmatrix} V^T. \tag{55}$$

Equation 54 is the key practical formula for the optimal solution $\mathbf{c}$ obtained by D-MORPH regression. This solution $\mathbf{c}_\infty$ is unique in $\mathcal{M}$ corresponding to the global minimum of the cost function. The new solution $\mathbf{c}_\infty$ given by D-MORPH regression is simply a linear combination of the elements of $\mathbf{c}$ obtained by least-squares regression (i.e., $A^+\mathbf{d}$).

3.2.2 Construction of the cost function

The solution of Eq. 44 satisfying the hierarchical orthogonality condition can be determined by constructing a proper cost function. The cost function may be deduced as follows.

The 1st order component function, $f_i(x_i)$, is orthogonal to the zeroth order component function, f_0, i.e.,

$$\int f_0 f_i(x_i) w_i(x_i)\mathrm{d}x_i = f_0 \int f_i(x_i) w_i(x_i)\mathrm{d}x_i = 0, \quad (i = 1, 2, \ldots, n). \tag{56}$$

Since f_0 may be nonzero, we require

$$\int f_i(x_i) w_i(x_i)\mathrm{d}x_i = 0, \quad (i = 1, 2, \ldots, n). \tag{57}$$

When $f_i(x_i)$ is represented as a linear combination of basis functions $r_j(x_i)(j = q+1, \ldots, q+k)$,[3] we have

$$\int \sum_{j=q+1}^{q+k} c_j r_j(x_i) w_i(x_i)\mathrm{d}x_i \approx \sum_{j=q+1}^{q+k} c_j \left(\sum_{s=1}^{N} r_j(x_i^{(s)})/N \right) = 0. \tag{58}$$

[3] For optimal orthonormal polynomials, k is often chosen as 2 or 3; for B-splines, k depends on the number of knots used.

Equation 58 can be written as

$$\left(\sum_{s=1}^{N} r_{q+1}(x_i^{(s)})/N \ \sum_{s=1}^{N} r_{q+2}(x_i^{(s)})/N \ \cdots \ \sum_{s=1}^{N} r_{q+k}(x_i^{(s)})/N \right) \begin{pmatrix} c_{q+1} \\ c_{q+2} \\ \vdots \\ c_{q+k} \end{pmatrix} = 0 \tag{59}$$

or in vector form

$$\mathbf{Sr}(x_i)^T \mathbf{c}^i = 0, \quad (i = 1, 2, \ldots, n). \tag{60}$$

This is a scalar equation, and the corresponding cost function for $f_i(x_i)$ is set to be

$$\mathcal{K}^i = \frac{1}{2}(\mathbf{c}^i)^T \mathbf{Sr}(x_i)\mathbf{Sr}(x_i)^T \mathbf{c}^i = \frac{1}{2}(\mathbf{c}^i)^T B^i \mathbf{c}^i, \quad (i = 1, 2, \ldots, n) \tag{61}$$

where B^i is a $k \times k$ symmetric and non-negative definite matrix. Therefore, $\mathcal{K}^i \geq 0$ with the minimum value being zero. $\mathcal{K}^i$ is zero if and only if $\mathbf{Sr}(x_i)^T \mathbf{c}^i$ is zero, i.e., Eq. 60 (consequently, Eq. 57) is satisfied. When optimal orthonormal polynomials are used as $r_j(x_i)$, then all of the sums $\sum_{s=1}^{N} r_{q+j}(x_i^{(s)})/N (j = 1, 2, 3)$ are zero (see Eqs. 32–34). In this circumstance B^i is a null matrix, which implies that there is no need for further restriction on the expansion coefficients for $f_i(x_i)$ upon using optimal orthonormal polynomials.

The 2nd order component function $f_{ij}(x_i, x_j)$ is required to be orthogonal to f_0 and the 1st order component functions, $f_i(x_i)$ and $f_j(x_j)$. This can be achieved by setting $f_{ij}(x_i, x_j)$ to be orthogonal to all the basis functions used in f_0 (its basis is 1), $f_i(x_i)$ and $f_j(x_j)$. Since $f_{ij}(x_i, x_j)$ is orthogonal to all the basis functions, it must be orthogonal to any linear combination of these basis functions, and consequently orthogonal to f_0, $f_i(x_i)$ and $f_j(x_j)$.

Let

$$f_i(x_i) = \sum_{l=1}^{k} c_l^i r_l^i(x_i), \tag{62}$$

$$f_j(x_j) = \sum_{l=1}^{k} c_l^j r_l^j(x_j), \tag{63}$$

$$f_{ij}(x_i, x_j) = \sum_{l=1}^{k} c_l^{(ij)i} r_l^i(x_i) + \sum_{l=1}^{k} c_l^{(ij)j} r_l^j(x_j) + \sum_{p=1}^{l'} \sum_{q=1}^{l'} c_{pq}^{(0)ij} r_p^i(x_i) r_q^j(x_j). \tag{64}$$

The orthogonality between $f_{ij}(x_i, x_j)$ and f_0 is given by

$$\int 1\left(\sum_{l=1}^{k} c_l^{(ij)i} r_l^i(x_i) + \sum_{l=1}^{k} c_l^{(ij)j} r_l^j(x_j) + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij} r_p^i(x_i) r_q^j(x_j)\right)$$
$$w_{ij}(x_i, x_j)\mathrm{d}x_i\mathrm{d}x_j$$
$$\approx \sum_{l=1}^{k} c_l^{(ij)i}\left(\sum_{s=1}^{N} r_l^i(x_i^{(s)})/N\right) + \sum_{l=1}^{k} c_l^{(ij)j}\left(\sum_{s=1}^{N} r_l^j(x_j^{(s)})/N\right)$$
$$+ \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij}\left(\sum_{s=1}^{N} r_p^i(x_i^{(s)}) r_q^j(x_j^{(s)})/N\right)$$
$$= \mathbf{Sr}_0(x_i, x_j)^T \mathbf{c}^{ij} = 0, \tag{65}$$

where $\mathbf{c}^{ij}$ is a $t_{ij}(= 2k + (l')^2)$-dimensional vector consisting of all expansion coefficients for $f_{ij}(x_i, x_j)$.

The orthogonality between $f_{ij}(x_i, x_j)$ and the basis $r_v^i(x_i)$ is given by

$$\int r_v^i(x_i)\left(\sum_{l=1}^{k} c_l^{(ij)i} r_l^i(x_i) + \sum_{l=1}^{k} c_l^{(ij)j} r_l^j(x_j) + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij} r_p^i(x_i) r_q^j(x_j)\right)$$
$$w_{ij}(x_i, x_j)\mathrm{d}x_i\mathrm{d}x_j = \sum_{l=1}^{k} c_l^{(ij)i}\langle r_v^i(x_i), r_l^i(x_i)\rangle$$
$$+ \sum_{l=1}^{k} c_l^{(ij)j}\langle r_v^i(x_i), r_l^j(x_j)\rangle + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij}\langle r_v^i(x_i), r_p^i(x_i) r_q^j(x_j)\rangle$$
$$\approx \sum_{l=1}^{k} c_l^{(ij)i}\left(\sum_{s=1}^{N} r_v^i(x_i^{(s)}) r_l^i(x_i^{(s)})/N\right) + \sum_{l=1}^{k} c_l^{(ij)j}\left(\sum_{s=1}^{N} r_v^i(x_i^{(s)}) r_l^j(x_j^{(s)})/N\right)$$
$$+ \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij}\left(\sum_{s=1}^{N} r_v^i(x_i^{(s)}) r_p^i(x_i^{(s)}) r_q^j(x_j^{(s)})/N\right)$$
$$= \mathbf{Sr}_{iv}(x_i, x_j)^T \mathbf{c}^{ij} = 0, \quad (v = 1, 2, \ldots, k) \tag{66}$$

where the elements of $\mathbf{Sr}_{iv}(x_i, x_j)^T$ are the estimates of the inner products of $r_v^i(x_i)$ and all the basis functions used by $f_{ij}(x_i, x_j)$. The orthogonality between $r_v^j(x_j)$ and $f_{ij}(x_i, x_j)$ can be treated similarly, which gives

$$\mathbf{Sr}_{jv}(x_i, x_j)^T \mathbf{c}^{ij} = 0, \quad (v = 1, 2, \ldots, k) \tag{67}$$

All together there are $2k + 1$ equations in Eqs. 65–67, which can be represented in matrix form

$$\mathbf{Sr}(x_i, x_j)^T \mathbf{c}^{ij} = \mathbf{0}, \tag{68}$$

where $\mathbf{Sr}(x_i, x_j)^T$ is a $(2k + 1) \times t_{ij}$ matrix, and $\mathbf{0}$ is a $(2k + 1)$-dimensional null vector.

The cost function for the orthogonality between $f_{ij}(x_i, x_j)$ and $f_0,\ f_i(x_i),\ f_j(x_j)$ is specified as

$$\begin{aligned}\mathcal{K}^{ij} &= \frac{1}{2}(\mathbf{c}^{ij})^T \mathbf{Sr}(x_i, x_j)\mathbf{Sr}(x_i, x_j)^T \mathbf{c}^{ij}\\ &= \frac{1}{2}(\mathbf{c}^{ij})^T B^{ij}\mathbf{c}^{ij}, \quad (i < j = 1, 2, \ldots, n)\end{aligned} \tag{69}$$

where B^{ij} is a $t_{ij} \times t_{ij}$ symmetric, non-negative definite matrix. Therefore, $\mathcal{K}^{ij} \geq 0$ with a minimum value of zero which occurs if and only if $\mathbf{Sr}(x_i, x_j)^T \mathbf{c}^{ij}$ is a null vector, i.e., $f_{ij}(x_i, x_j)$ is orthogonal to f_0 and all $r_v^i(x_i)$ and $r_v^j(x_j)$.

A similar treatment can be made for $f_{ijk}(x_i, x_j, x_k)$, and the corresponding cost function

$$\mathcal{K}^{ijk} = \frac{1}{2}(\mathbf{c}^{ijk})^T B^{ijk}\mathbf{c}^{ijk}, \quad (i < j < k = 1, 2, \ldots, n) \tag{70}$$

can be constructed. If the 3rd order HDMR expansion is used, the total cost function is set to be

$$\begin{aligned}\mathcal{K} &= \sum_{i=1}^{n}\mathcal{K}^i + \sum_{1\leq i<j\leq n}\mathcal{K}^{ij} + \sum_{1\leq i<j<k\leq n}\mathcal{K}^{ijk}\\ &= \frac{1}{2}\mathbf{c}^T B\mathbf{c}\end{aligned} \tag{71}$$

where $\mathbf{c}$ consists of all the unknown coefficients in Eq. 35, and

$$B = \begin{pmatrix} B^1 &&&&&&&& \\ & \ddots &&&&&&& \\ && B^n &&&&&& \\ &&& B^{12} &&&&& \\ &&&& \ddots &&&& \\ &&&&& B^{(n-1)n} &&& \\ &&&&&& B^{123} && \\ &&&&&&& \ddots & \\ &&&&&&&& B^{(n-2)(n-1)n} \end{pmatrix} \tag{72}$$

is a non-negative definite matrix. Therefore, $\mathcal{K} \geq 0$ and its minimum value is zero which implies the hierarchical orthogonality of the component functions. All B^{ij} and B^{ijk} are submatrices of $\mathbf{\Phi}^T\mathbf{\Phi}/N$ and can be obtained from it.

Importantly, the extended basis method and D-MORPH regression with the cost function defined above also can be applied to the cases with independent variables. For example, suppose that x_i and x_j are independent and optimal orthonormal polynomial bases are used. Then we have

$$\begin{aligned}\langle r_p^i(x_i), r_q^j(x_j)\rangle &= \int r_p^i(x_i) r_q^j(x_j) w_{ij}(x_i, x_j) \mathrm{d}x_i \mathrm{d}x_j \\ &= \int r_p^i(x_i) w_i(x_i) \mathrm{d}x_i \int r_q^j(x_j) w_j(x_j) \mathrm{d}x_j = 0. \end{aligned} \tag{73}$$

and

$$\begin{aligned}\langle r_v^i(x_i), r_p^i(x_i) r_q^j(x_j)\rangle &= \int r_v^i(x_i) r_p^i(x_i) r_q^j(x_j) w_{ij}(x_i, x_j) \mathrm{d}x_i \mathrm{d}x_j \\ &= \int r_v^i(x_i) r_p^i(x_i) w_i(x_i) \mathrm{d}x_i \int r_q^j(x_j) w_j(x_j) \mathrm{d}x_j = 0. \end{aligned} \tag{74}$$

Substituting Eqs. 73, 74 into Eq. 66 yields

$$\mathbf{Sr}_{iv}(x_i, x_j)^T \mathbf{c}^{ij} = c_v^{(ij)i} \langle r_v^i(x_i), r_v^i(x_i)\rangle = c_v^{(ij)i} = 0, \quad (v = 1, 2, \ldots, k). \tag{75}$$

Similarly we have $c_v^{(ij)j} = 0 (v = 1, 2, \ldots, k)$. Then Eq. 64 becomes

$$f_{ij}(x_i, x_j) = \sum_{p=1}^{l'} \sum_{q=1}^{l'} c_{pq}^{(0)ij} r_p^i(x_i) r_q^j(x_j), \tag{76}$$

i.e., the extended basis reduces to the non-extended basis. Moreover, distinct $r_p^i(x_i) r_q^j(x_j)$'s are orthogonal for independent x_i and x_j. The current algorithm reduces to the prior procedure for treating independent variables [6,13]. Therefore, no matter whether some or all of the variables are independent and/or correlated, the D-MORPH regression method with extended bases applies.

4 Illustration

Consider the following nonlinear function with three variables $\mathbf{x} = (x_1, x_2, x_3)$

$$f(\mathbf{x}) = g_1(x_1, x_2) + g_2(x_2) + g_3(x_3), \tag{77}$$

which will be used here for illustrating the general treatment in Sects. 2 and 3. The function is specified by

$$\begin{aligned} g_1(x_1, x_2) &= g_{1a}(x_1)g_{1b}(x_2) \\ &= [a_1(x_1 - \mu_1) + a_0][b_1(x_2 - \mu_2) + b_0], \quad (78) \\ g_2(x_2) &= c_2(x_2 - \mu_2)^2 + c_1(x_2 - \mu_2) + c_0, \quad (79) \\ g_3(x_3) &= d_3(x_3 - \mu_3)^3 + d_2(x_3 - \mu_3)^2 + d_1(x_3 - \mu_3) + d_0 \quad (80) \end{aligned}$$

with a multivariate normal distribution

$$w(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right), \quad (81)$$

where $\mu = (\mu_1, \mu_2, \mu_3)$ is the expected value of $\mathbf{x}$, $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{x}$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & 0 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}, \quad (82)$$

i.e., x_1 and x_2 are correlated, but x_3 is independent. The structure of the model in Eqs. 77–82 permits an analytical solution of the unique HDMR component functions below for comparison with the numerical approximations.

As the unique HDMR component functions are coupled together in the formulas given in Eqs. 26–28, generally a closed form for each component function with correlated variables cannot be obtained as was the case in Eqs. 5–7 for independent variables. However, approximate formulas may be constructed by using the following procedure:

1. Set a maximum order for the truncated HDMR expansion. For instance, if a 2nd order expansion is used, the higher order component functions are set to zero.
2. Approximate the functions $h_i(x_i)$, $h_{ij}(x_i, x_j)$ by an expansion in some suitable basis functions. The determination of $h_i(x_i)$ and $h_{ij}(x_i, x_j)$ then reduces to identifying the constant expansion coefficients.
3. Apply the vanishing condition, Eq. 19 or Eq. 21, to all the component functions. This procedure yields a set of linear algebraic equations, which may be solved for the constant coefficients.

As the model in Eq. 77 consists of polynomials with order less than 3, its HDMR expansion only contains zeroth, 1st and 2nd order HDMR component functions. Using the procedures given above, exact analytical formulas of the unique HDMR component functions for $f(\mathbf{x})$ can be obtained. To save space, the details are not given here;

only the resultant component functions are provided.

$$f_0 = a_1 b_1 \rho_{12} \sigma_1 \sigma_2 + a_0 b_0 + c_2 \sigma_2^2 + c_0 + d_2 \sigma_3^2 + d_0, \tag{83}$$

$$\begin{aligned} f_1(x_1) = & a_1 b_1 \frac{\sigma_2}{\sigma_1} \frac{\rho_{12}}{\rho_{12}^2 + 1} (x_1 - \mu_1)^2 + a_1 b_0 (x_1 - \mu_1) \\ & - a_1 b_1 \sigma_1 \sigma_2 \frac{\rho_{12}}{\rho_{12}^2 + 1}, \end{aligned} \tag{84}$$

$$\begin{aligned} f_2(x_2) = & \left[a_1 b_1 \frac{\sigma_1}{\sigma_2} \frac{\rho_{12}}{\rho_{12}^2 + 1} + c_2 \right] (x_2 - \mu_2)^2 \\ & + (a_0 b_1 + c_1)(x_2 - \mu_2) - a_1 b_1 \sigma_1 \sigma_2 \frac{\rho_{12}}{\rho_{12}^2 + 1} - c_2 \sigma_2^2, \end{aligned} \tag{85}$$

$$f_3(x_3) = d_3 (x_3 - \mu_3)^3 + d_2 (x_3 - \mu_3)^2 + d_1 (x_3 - \mu_3) - d_2 \sigma_3^2, \tag{86}$$

$$\begin{aligned} f_{12}(x_1, x_2) = & - a_1 b_1 \frac{\sigma_2}{\sigma_1} \frac{\rho_{12}}{\rho_{12}^2 + 1} (x_1 - \mu_1)^2 + a_1 b_1 (x_1 - \mu_1)(x_2 - \mu_2) \\ & - a_1 b_1 \frac{\sigma_1}{\sigma_2} \frac{\rho_{12}}{\rho_{12}^2 + 1} (x_2 - \mu_2)^2 - a_1 b_1 \rho_{12} \sigma_1 \sigma_2 \frac{\rho_{12}^2 - 1}{\rho_{12}^2 + 1}, \end{aligned} \tag{87}$$

$$f_{13}(x_1, x_3) = 0, \tag{88}$$

$$f_{23}(x_2, x_3) = 0. \tag{89}$$

All the component functions satisfy the vanishing condition Eq. 19 (or Eq. 21) and are hierarchically orthogonal. From the above formulas we find that:

1. The sum of all the component functions is exactly equal to $f(\mathbf{x})$, i.e., the second order HDMR expansion is an exact representation for $f(\mathbf{x})$.
2. The component functions reflect not only the deterministic relation, Eq. 77, but also the multivariate normal distribution. For example, $f(\mathbf{x})$ is a linear function of $x_1 - \mu_1$, but $f_1(x_1)$ is quadratic in $x_1 - \mu_1$ because $f(\mathbf{x})$ is a function of $(x_1 - \mu_1)(x_2 - \mu_2)$ and x_1 is correlated with x_2.
3. x_3 is independent, and the formula for $f_3(x_3)$ is the same as that given by Eq. 6 for independent variables.

Since the exact analytical formulas of the component functions are known in this example, below we will test the effectiveness of the numerical method based on D-MORPH regression with extended bases.

4.1 Results for error free data

Six hundred points of $\mathbf{x}$ were generated according to the multivariate normal distribution with $\sigma_1 = \sigma_2 = 0.2$, $\sigma_3 = 0.18$ and $\rho_{12} = 0.6$, and the corresponding values of $f(\mathbf{x})$ were calculated for $a_0 = 1, a_1 = 2, b_0 = 2, b_1 = 3, c_0 = 3, c_1 = 1, c_2 = 2, d_0 = 1, d_1 = 2, d_2 = 2, d_3 = 3$ (referred to as Data set 1). The first three hundred points were used as training data, and the others were used for testing.

Table 1 Estimates of the correlation coefficients ρ_{ij} from different sample sizes

Data size	ρ_{12}	ρ_{13}	ρ_{23}
300	0.5668	−0.1145	−0.0128
600	0.5647	−0.0994	0.0067
1000	0.5765	−0.0679	0.0126
2000	0.5990	−0.0145	0.0276
10000	0.5955	0.0099	0.0190
30000	0.6012	−0.0066	0.0079

4.1.1 Model with 54 unknown parameters

Equations 83–89 show that component functions higher than second order do not exist. Thus, the second order HDMR expansion with up to 3rd order optimal orthonormal polynomial bases was employed to represent $f(\mathbf{x})$

$$
\begin{aligned}
f(\mathbf{x}) &= f_0 + \sum_{i=1}^{3} f_i(x_i) + \sum_{1\le i<j\le 3} f_{ij}(x_i, x_j) \\
&= f_0 + \sum_{i=1}^{3}\sum_{r=1}^{3} \alpha_r^{(0)i}\varphi_r^i(x_i) + \sum_{1\le i<j\le 3}\sum_{r=1}^{3}\left[\alpha_r^{(ij)i}\varphi_r^i(x_i) + \alpha_r^{(ij)j}\varphi_r^j(x_j)\right] \\
&\quad + \sum_{p=1}^{3}\sum_{q=1}^{3}\beta_{pq}^{(0)ij}\varphi_p^i(x_i)\varphi_q^j(x_j).
\end{aligned}
\tag{90}
$$

The model contains all $f_i(x_i)(i = 1, 2, 3)$ and $f_{ij}(x_i, x_j)(i < j = 1, 2, 3)$. Collectively there are 54 unknown parameters ($\alpha_r^{(0)i}$, $\alpha_r^{(ij)i}$, $\alpha_r^{(ij)j}$ and $\beta_{pq}^{(0)ij}$).

Since optimal orthonormal polynomials are used, no further restriction is needed for the coefficients $\alpha_r^{(0)i}$ and the B^i's are null matrices. Care is needed in the construction of B^{ij} for the cost function. For example, we know that $f_{13}(x_1, x_3) = f_{23}(x_2, x_3) = 0$, and a zero function is strictly orthogonal to any other function. Including f_{13}, f_{23} in the regression model may result in a large bias. As x_3 is independent, the correlation coefficients satisfy $\rho_{13} = \rho_{23} = 0$. However, the estimates of the correlation coefficients based on a small sample size N

$$
\rho_{ij} = \frac{\sum_{s=1}^{N}\left[(x_i^{(s)} - \bar{x}_i)(x_j^{(s)} - \bar{x}_j)\right]}{\sqrt{\sum_{s=1}^{N}(x_i^{(s)} - \bar{x}_i)^2}\sqrt{\sum_{s=1}^{N}(x_j^{(s)} - \bar{x}_j)^2}},
\tag{91}
$$

where $\bar{x}_i$ is the average value of $x_i^{(s)}$, can produce large errors. Table 1 gives the estimates of ρ_{ij} with different sample sizes.

When 300 samples are used, the results significantly differ from $\rho_{12} = 0.6$, $\rho_{13} = \rho_{23} = 0$, and the estimates of the inner products $\langle r_p^i(x_i), r_q^3(x_3)\rangle(i = 1, 2)$ used in the construction of B^{ij}, and consequently the resultant $f_{13}(x_1, x_3)$ and $f_{23}(x_2, x_3)$ will also significantly differ from zero. In order to test the new method, in the construction of the cost function we simply set $\langle r_p^i(x_i), r_q^3(x_3)\rangle = 0\,(i = 1, 2)$.

The comparison between the analytical solution and the solution obtained by D-MORPH regression with extended bases is given in Fig. 1. The figure shows that using the model with 54 unknown parameters and the cost function given above, the component functions $f_i(x_i)$ obtained by D-MORPH regression with extended bases coincide very well with the analytical solution. The approximate results for $f_{12}(x_1, x_2)$ obtained by D-MORPH regression with extended bases does not coincide with the analytical solution for an arbitrary point (x_1, x_2). However, the numerical and analytical results do agree for those pairs of points which satisfy the correlated pdf, because the estimated $f_{12}(x_1, x_2)$ is obtained only from the data generated by the pdf. To compare the numerical and analytical solutions, the truth plot for $f_{12}(x_1, x_2)$ is given at the sampled points according to the pdf. Some points are not exactly located on the truth plot line. This was caused by the limited sample size (300 points). The resultant terms $\mathbf{Sr}_0(x_1, x_2)$, $\mathbf{Sr}_{1v}(x_1, x_2)$ and $\mathbf{Sr}_{2v}(x_1, x_2)$ have errors just like the determination of ρ_{ij}. Figure 1 also shows that $f_{13}(x_1, x_3)$ and $f_{23}(x_1, x_3)$ obtained by the numerical method at the sampled points are practically zero. Therefore, their values are simply shown according to their sample order. Figure 1 demonstrates that the component functions are unique, and the new numerical method accurately determines them.

The accuracy of least-squares regression given by Eq. 47 and D-MORPH regression with extended bases is given in Table 2 and Fig. 2. Since the model, Eq. 90, contains f_{13} and f_{23} which do not exist in the true system, the least-squares solution (Eq. 47) is far from the true system which causes a large error. In contrast, D-MORPH regression with extended bases accurately finds the true system.

4.1.2 Model with 24 unknown parameters

In practice, we do not know which component functions, such as $f_{13}(x_1, x_3)$ and $f_{23}(x_1, x_3)$, are zero in advance. Therefore, *a priori* setting $\langle r_p^i(x_i), r_q^3(x_3)\rangle = 0\,(i = 1, 2)$ is not proper. To deal with this situation, we first use a statistical test (F-test) to identify the significant HDMR component functions from the training data [13,17]. In the present example, this gives

$$
\begin{aligned}
f(\mathbf{x}) &= f_0 + \sum_{i=1}^{3} f_i(x_i) + f_{12}(x_1, x_2) \\
&= f_0 + \sum_{i=1}^{3}\sum_{r=1}^{3} \alpha_r^{(0)i}\varphi_r^i(x_i) + \sum_{r=1}^{3}\left[\alpha_r^{(12)1}\varphi_r^1(x_1) + \alpha_r^{(12)2}\varphi_r^2(x_2)\right] \\
&\quad + \sum_{p=1}^{3}\sum_{q=1}^{3} \beta_{pq}^{(0)12}\varphi_p^1(x_1)\varphi_q^2(x_2)
\end{aligned}
\tag{92}
$$

with 24 unknown parameters ($\alpha_r^{(0)i}$, $\alpha_r^{(12)1}$, $\alpha_r^{(12)2}$ and $\beta_{pq}^{(0)12}$). Note that the model in Eq. 92 does not contain f_{13} and f_{23}, as they were identified as insignificant by

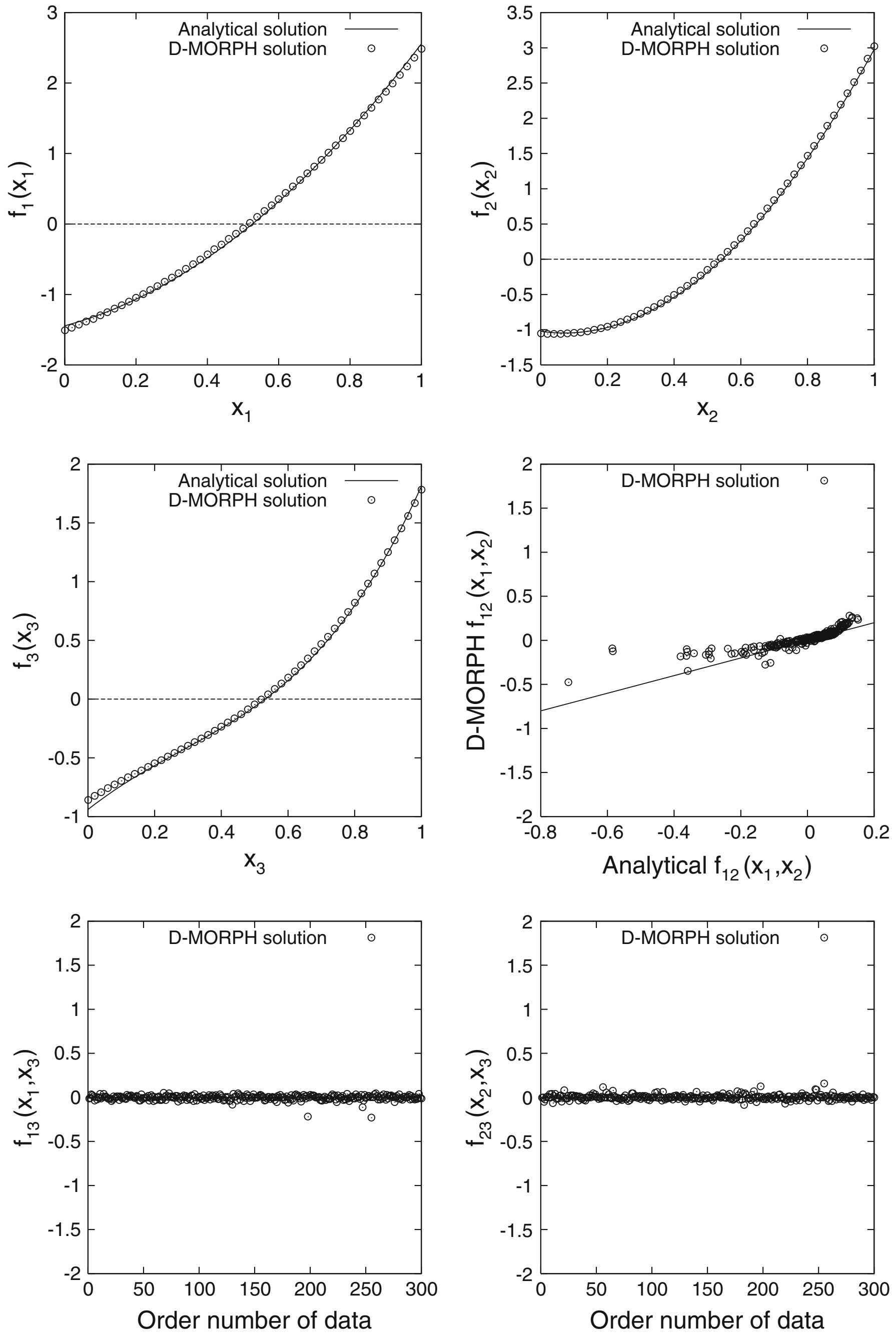

Fig. 1 Comparison of $f_i(x_i)$, $f_{ij}(x_i, x_j)$ obtained analytically and numerically by D-MORPH regression with extended bases for the model having 54 unknown parameters and Data set 1

Table 2 Accuracy of least-squares regression and D-MORPH regression with extended bases for the model having 54 unknown parameters and Data set 1

Data	Least-squares regression		D-MORPH regression	
	Ave. abs. err.	Ave. rel. err.	Ave. abs. err.	Ave. rel. err.
Training	0.410	0.069	0.033	0.006
Testing	0.431	0.074	0.045	0.007

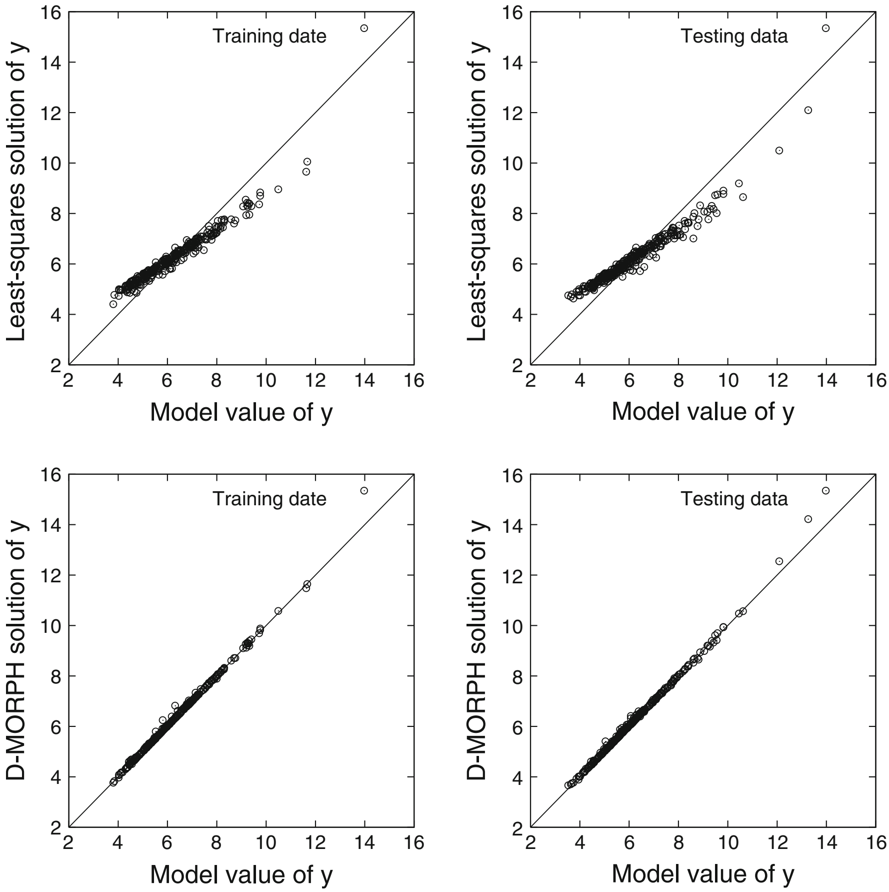

Fig. 2 Truth plots of training and testing data for least-squares regression and D-MORPH regression with extended bases for the model having 54 parameters and Data set 1

the F-test procedure. Both least-squares regression given by Eq. 47 and D-MORPH regression with extended bases were used. For the model with 24 unknown parameters the accuracy for the two methods is the same. The results are given in Table 3 and Fig. 3.

Table 3 Accuracy of least-squares regression and D-MORPH regression with extended bases for the model having 24 unknown parameters and Data set 1

Data	Least-squares regression		D-MORPH regression	
	Ave. abs. err.	Ave. rel. err.	Ave. abs. err.	Ave. rel. err.
Training	0.035	0.006	0.035	0.006
Testing	0.045	0.007	0.045	0.007

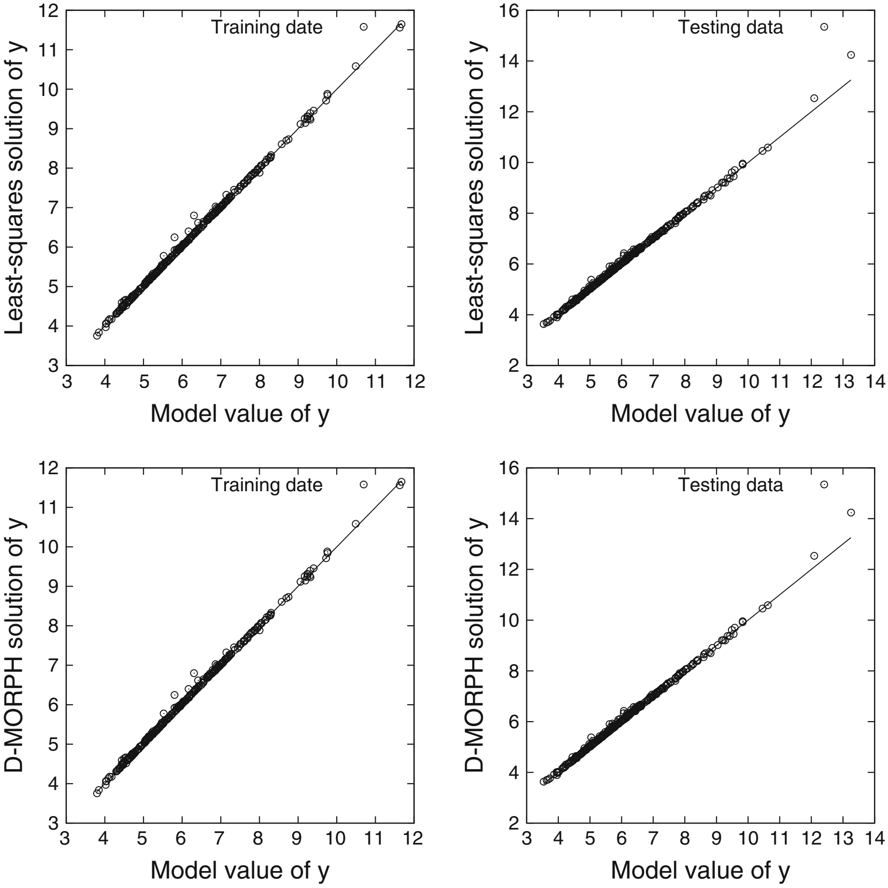

Fig. 3 Truth plots of training and testing data obtained by least-squares regression and D-MORPH regression with extended bases for the model having 24 unknown parameters and Data set 1

Even though the accuracy for the two methods is the same, the resultant component functions are distinct. $f_1(x_1)$ and $f_2(x_2)$ are orthogonal to $f_{12}(x_1, x_2)$ for D-MORPH regression with extended bases, but they are not orthogonal for least-squares regression. Figures 4 and 5 show that the component functions obtained by D-MORPH regression with extended bases accurately coincide with the analytical

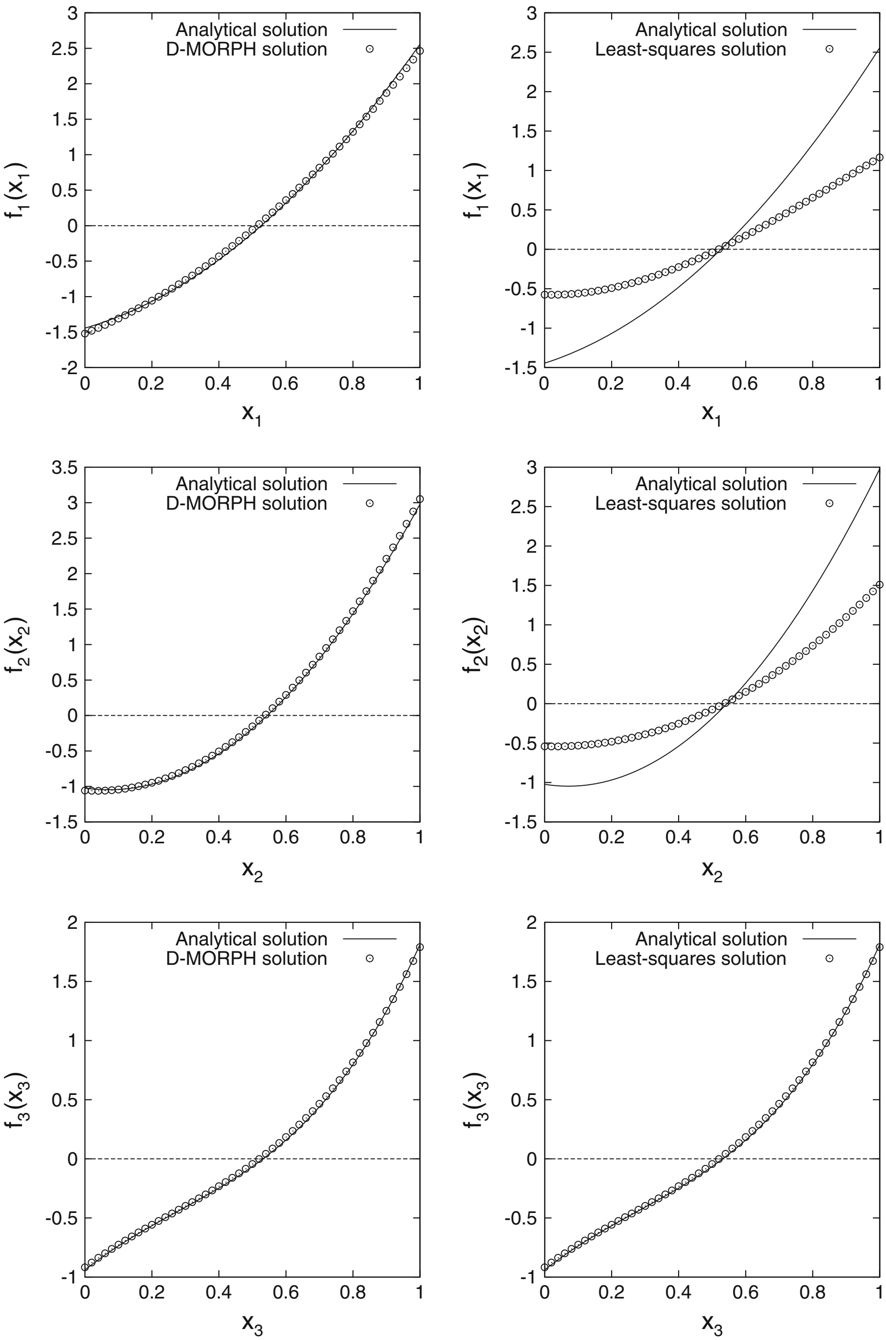

Fig. 4 Comparison of $f_i(x_i)$ $(i = 1, 2, 3)$ obtained analytically and numerically by least-squares and D-MORPH regressions with extended bases for the model having 24 unknown parameters and Data set 1

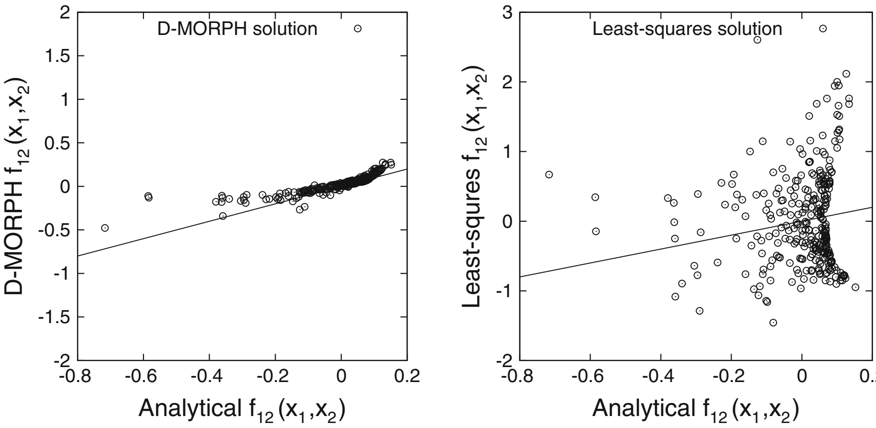

Fig. 5 Comparison of $f_{12}(x_1, x_2)$ obtained analytically and numerically by least-squares and D-MORPH regressions with extended bases for the model having 24 unknown parameters and Data set 1

Table 4 Orthogonality of the component functions obtained by least-squares regression and D-MORPH regression with extended bases for the model having 24 unknown parameters and Data set 1

Data	Least-squares regression			D-MORPH regression		
	Cost	$\langle f_1, f_{12}\rangle$	$\langle f_2, f_{12}\rangle$	Cost	$\langle f_1, f_{12}\rangle$	$\langle f_2, f_{12}\rangle$
Training	0.209	0.215	0.251	0.019	0.000	0.000
Testing	–	0.231	0.273	–	0.003	0.015

solutions, while those (except of $f_3(x_3)$ with independent variable x_3) obtained by least-squares regression do not.

The value of the cost function and the estimates of the inner products $\langle f_1(x_1), f_{12}(x_1, x_2)\rangle$, $\langle f_2(x_2), f_{12}(x_1, x_2)\rangle$ given in Table 4 also show the difference. Having unique and accurate HDMR component functions is essential for the reliable performance of global sensitivity analysis.

4.2 Results for data with random errors

To test the influence of random errors on the determination of the HDMR component functions by D-MORPH regression with extended bases, random noise with a signal to noise ratio $(\mathrm{Var}(f(\mathbf{x}))/\sigma^2) = 100$ was added to $f(\mathbf{x})$ to generate another set of data (referred to as Data set 2). A comparison of Data sets 1 and 2 is given in Fig. 6.

Using the model with 24 unknown parameters, least-squares regression and D-MORPH regression with extended bases were used to treat the data. The fitting accuracy is given in Table 5 and Fig. 7.

Similar to the error free data case, the accuracy for the two methods are the same, but the resultant HDMR component functions are distinct. Least-squares regression does not give the correct component functions. Comparison of the component functions

Fig. 6 Comparison of Data sets 1 and 2

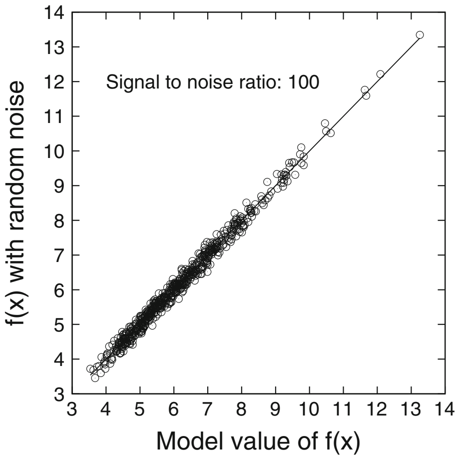

Table 5 Accuracy of least-squares regression and D-MORPH regression with extended bases for the model having 24 unknown parameters and Data set 2

Data	Least-squares regression		D-MORPH regression	
	Ave. abs. err.	Ave. rel. err.	Ave. abs. err.	Ave. rel. err.
Training	0.131	0.022	0.131	0.022
Testing	0.133	0.022	0.133	0.022

obtained analytically and numerically by D-MORPH regression with extended bases is given in Fig. 8, which shows that the random error did not have a significant influence on the solution (i.e., compare Fig. 8 with the plots in the first column of Figs. 4 and 5) and the resultant component functions still accurately coincide with the analytical solutions. This stability with respect to noise is important for global sensitivity analysis based on component functions obtained from laboratory/field data containing random errors.

5 Concluding remarks

Global sensitivity analysis is an important application of HDMR. In this regard, we recently introduced a new unified global sensitivity analysis framework for systems with independent and/or correlated variables referred to as Structural and Correlative Sensitivity Analysis (SCSA) [15]. The precondition for reliable application of SCSA is the generation of correct HDMR component functions. A high quality input–output map fitting the training data is a necessary, but not sufficient, condition for a reliable subsequent sensitivity analysis assessment. Therefore, the uniqueness of the HDMR component functions for independent and/or correlated variables is crucial for obtaining a physically meaningful global sensitivity analysis. Under the relaxed vanishing condition or hierarchical orthogonality condition, the general formulas for the unique

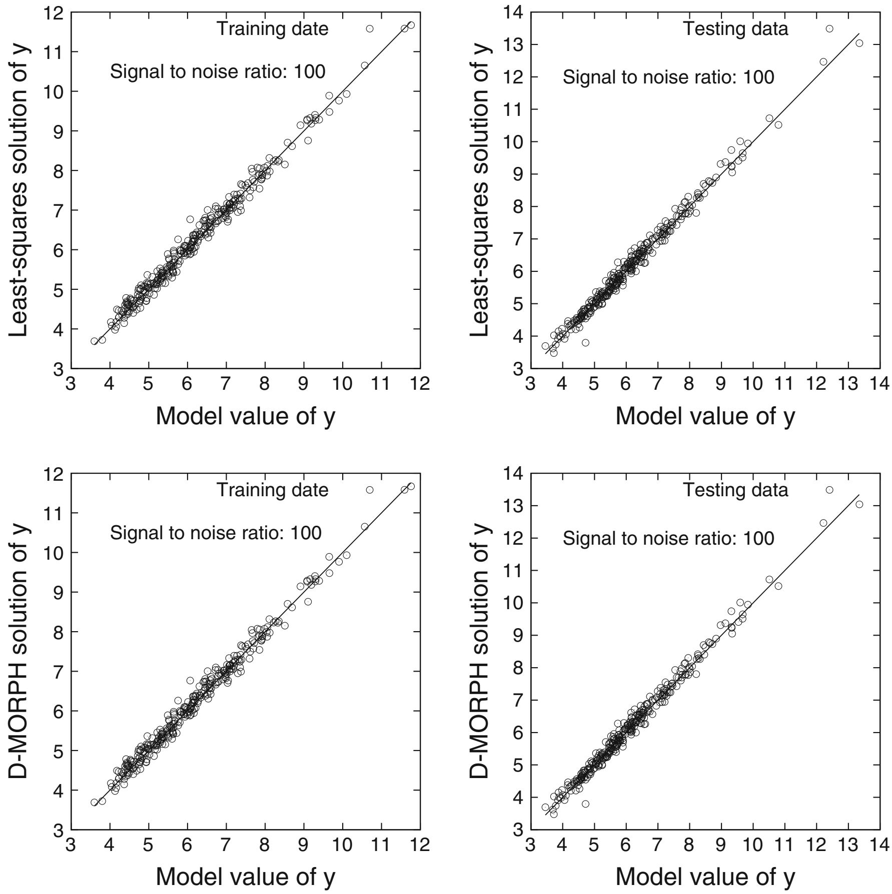

Fig. 7 Truth plots of training and testing data for least-squares regression and D-MORPH regression with extended bases for the model having 24 unknown parameters and Data set 2

HDMR component functions for independent and/or correlated variables are derived in this paper. The results reduce to the prior formulas used for independent variables [10,16] as a special case. As the general expression determining a component function with correlated variables contains all other component functions sharing the same common variables, generally a closed form for each component function may not be obtained. However, a novel and practical numerical method, combining extended bases and D-MORPH regression, can efficiently and accurately determine the unique HDMR component functions from available input–output data. This numerical method can be used for both independent and correlated variables. A reasonable level of random data error was found to not significantly influence the determination of the HDMR component functions. Thus, a unified framework is established for the unique HDMR decomposition of an n-variate function $f(\mathbf{x})$ with independent and/or correlated variables.

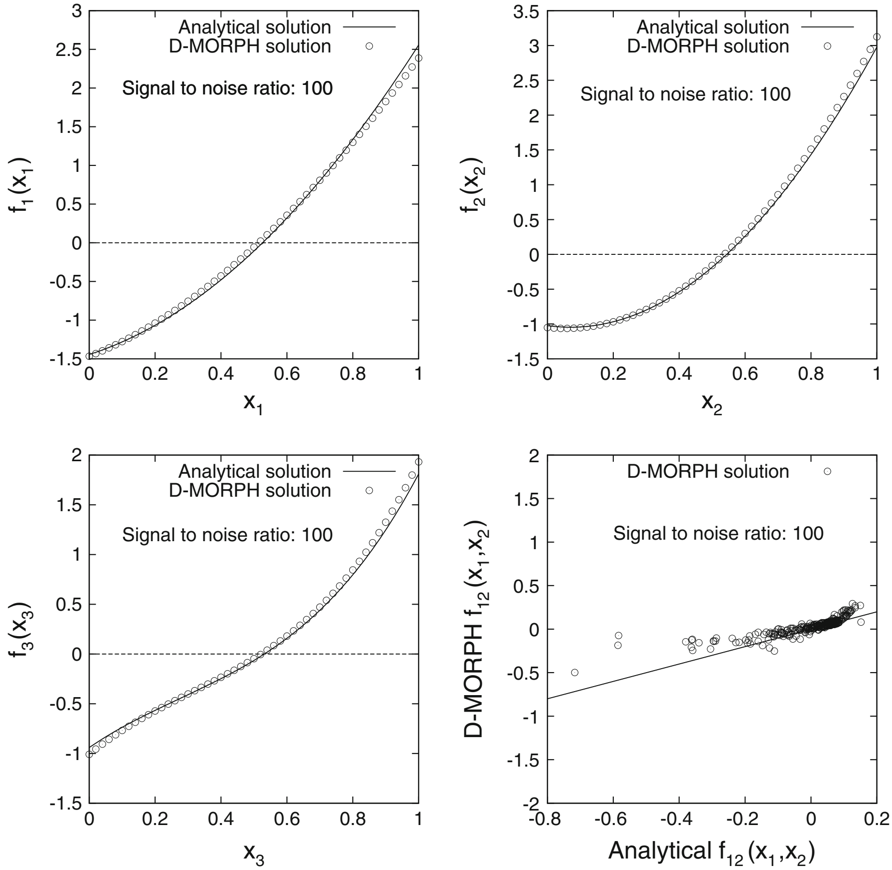

Fig. 8 Comparison of $f_i(x_i)$, $f_{12}(x_1, x_2)$ obtained analytically and numerically by D-MORPH regression with extended bases for the model having 24 unknown parameters and Data set 2

Acknowledgments Support for this work has been provided by the NSF and ONR.

References

1. H. Rabitz, Ö.F. Alis, J. Math. Chem. **25**, 197–233 (1999)
2. H. Rabitz, Ö.F. Alis, J. Shorter, K. Shim, Comput. Phys. Comm. **117**, 11–20 (1999)
3. J. Shorter, C.Op. Precis, H. Rabitz, J. Phys. Chem. A **103**(36), 7192–7198 (1999)
4. G. Li, C. Rosenthal, H. Rabitz, J. Phys. Chem. A **105**, 7765–7777 (2001)
5. G. Li, S.W. Wang, H. Rabitz, C. Rosenthal, J. Math. Chem. **30**, 1–30 (2001)
6. G. Li, S.W. Wang, H. Rabitz, J. Phys. Chem. **106**, 8721–8733 (2002)
7. G. Li, S.W. Wang, H. Rabitz, S. Wang, P.R. Jáffe, Chem. Eng. Sci. **57**, 4445–4460 (2002)
8. G. Li, H. Rabitz, S.W. Wang, P.G. Georgopoulos, J. Comput. Chem. **24**, 277–283 (2003)
9. G. Li, H. Rabitz, S.W. Wang, P.G. Georgopoulos, J. Comput. Chem. **24**, 647–656 (2003)
10. S.W. Wang, P.G. Georgopoulos, G. Li, H. Rabitz, J. Phys. Chem. A **107**, 4707–4718 (2003)
11. G. Li, H. Rabitz, J. Schoendorf, J. Comput. Chem. **25**, 1149–1156 (2004)
12. G. Li, H. Rabitz, J. Comput. Chem. **27**, 1112–1118 (2006)

13. G. Li, J.S. Hu, S.W. Wang, P.G. Georgopoulos, J. Schoendorf, H. Rabitz, J. Phys. Chem. A **110**, 2474–2485 (2006)
14. G. Li, H. Rabitz, J.S. Hu, Z. Chen, Y. Ju, J. Math. Chem. **43**(3), 1207–1232 (2008)
15. G. Li, H. Rabitz, P.E. Yelvington, O.O. Oluwole, F. Bacon, C.E. Kolb, J. Schoendorf, J. Phys. Chem. **114**, 6022–6032 (2010)
16. I.M. Sobol, Modeling and Computational Experiments **1**, 407–414 (1993)
17. T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer, New York, 2001)
18. J.E. Oakley, A. O'Hagan, J. R. Stat. Soc. **66**(Part 3), 751 (2004)
19. J.H. Friedman, Ann. Stat. **29**(5), 1189–1232 (2001)
20. G. Hooker, J. Comput. Graph. Stat. **16**(3), 709–732 (2007)
21. G. Li, H. Rabitz, J. Math. Chem. **48**(4), 1010–1035 (2010)
22. F. Deutsch, *Best approximation in inner product space* (Springer, New York, 2000)
23. C.R. Rao, S.K. Mitra, *Generalized inverse of matrix and its applications* (Willey, New York, 1971)
24. A. Rothman, T.-S. Ho, H. Rabitz, Phys. Rev. A **72**, 023416 (2005)
25. A. Rothman, T.-S. Ho, H. Rabitz, J. Chem. Phys. **123**, 134104 (2005)
26. A. Rothman, T.-S. Ho, H. Rabitz, Phys. Rev. A **73**, 053401 (2006)
27. N. Danielson, V. Beltrani, J. Dominy, H. Rabitz, manuscript in preparation
28. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical recipes in* FORTRAN—*The art of science computing*, 2nd edn. (Cambridge university press, New York, 1992), p. 51