Journal of Mathematical Chemistry Vol. 29, No. 2, 2001

Efficient implementation of high dimensional model representations

Ömer F. Alış [a] and Herschel Rabitz [b,*]

[a] *Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA*
[b] *Department of Chemistry, Princeton University, Princeton, NJ 08544, USA*

Received 08 December 2000

Physical models of various phenomena are often represented by a mathematical model where the output(s) of interest have a multivariate dependence on the inputs. Frequently, the underlying laws governing this dependence are not known and one has to interpolate the mathematical model from a finite number of output samples. Multivariate approximation is normally viewed as suffering from the curse of dimensionality as the number of sample points needed to learn the function to a sufficient accuracy increases exponentially with the dimensionality of the function. However, the outputs of most physical systems are mathematically well behaved and the scarcity of the data is usually compensated for by additional assumptions on the function (i.e., imposition of smoothness conditions or confinement to a specific function space). High dimensional model representations (HDMR) are a particular family of representations where each term in the representation reflects the individual or cooperative contributions of the inputs upon the output. The main assumption of this paper is that for most well defined physical systems the output can be approximated by the sum of these hierarchical functions whose dimensionality is much smaller than the dimensionality of the output. This ansatz can dramatically reduce the sampling effort in representing the multivariate function. HDMR has a variety of applications where an efficient representation of multivariate functions arise with scarce data. The formulation of HDMR in this paper assumes that the data is randomly scattered throughout the domain of the output. Under these conditions and the assumptions underlying the HDMR it is argued that *the number of samples needed for representation to a given tolerance is invariant to the dimensionality of the function*, thereby providing for a very efficient means to perform high dimensional interpolation. Selected applications of HDMR's are presented from sensitivity analysis and time-series analysis.

KEY WORDS: function representation, Monte Carlo integration

1. Introduction

The underlying map between the output of a physical system and its input variables is most often unknown or seriously lacking in *a priori* information. The characterization of the inputs is an important aspect of the modelling process and in this paper we will assume that it is known which inputs may potentially affect a certain output. Deducing

* Corresponding author.

0259-9791/01/0200-0127$19.50/0

the structure of an output depending on an n-dimensional space for $n \gg 1$ is an arduous task conventionally viewed as suffering from the *curse of dimensionality*. Conventional logic implies that the computational complexity of sampling the input-output map scales exponentially as s^n where s is a parameter specific to the problem and n is the relevant dimension. In applications, such maps arise from the underlying systems being either a computational model or an observational relationship. Severe difficulty in exploring the map occurs when a single output sample is expensive to attain, and then the efficiency of map representation and its need for sampling become critical. A related common problem where the curse of dimensionality arises is the integration of high dimensional functions [1]. The number of integrand evaluations increases exponentially with the dimension utilizing standard quadrature methods. The distribution of the samples used for high dimensional integration is crucial and it will be equally important in the multivariate approximation presented below. Monte Carlo and associated sampling schemes are the only practical means for performing high dimensional integrals. The difference between the sampling cost for high dimensional integration using a regular grid versus Monte Carlo integration is dramatic, as the former scales exponentially with n while the latter is dimension independent [2], assuming that the integrand is sufficently smooth.

One way to deal with the curse of dimensionality in approximating multivariate functions is to introduce dimension reduction techniques often inspired by a theorem of Kolmogorov [3] which states that a multivariate function defined on the unit cube $K^n = [0, 1]^n$ can be represented the following way:

$$f(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{2n+1} g\big(\lambda_1 \phi_q(x_1) + \cdots + \lambda_n \phi_q(x_n)\big), \tag{1}$$

i.e., any multivariate function of $x \equiv (x_1, \ldots, x_n)$ can be written as a linear superposition of univariate functions. Although the functions ϕ_q are continuous, they are highly non-smooth and their practical utility for approximation/interpolation has yet to be demonstrated [4]. Various approximation techniques resembling the above form have been utilized to represent a function $f(x)$ as a sum of superpositions of low-dimensional functions. Projection pursuit algorithms, multilayer perceptrons and radial basis networks are most commonly used as dimension reduction techniques [5–9]. Another representation which writes multivariate output as a superposition of functions of fewer variables is the function-space extension of the analysis of variance (ANOVA) decomposition commonly used in statistics to analyze the variance of a statistical quantity [10–12]. The ANOVA decomposition is a specific member of high dimensional model representations (HDMR) [13]. HDMR's assume the following exact form for the multivariate function:

$$f(x) \equiv f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots + f_{12\ldots n}(x_1, x_2, \ldots, x_n). \tag{2}$$

Here f_0 denotes the zeroth order effect which is a constant everywhere in the domain of $f(x)$. The function $f_i(x_i)$ gives the effect associated with the variable x_i acting independently, although generally nonlinearly, upon the output f. The second order function

$f_{ij}(x_i, x_j)$ describes the cooperative effects of the variables x_i and x_j and higher order terms reflect the cooperative effects of increasing numbers of variables acting together to impact upon f. The last term $f_{12\ldots n}(x_1, \ldots, x_n)$ gives any residual dependence of all the variables locked together in a cooperative way to influence the output f. For a well-defined problem with rationally chosen physical variables it is natural to expect that convergence will occur at relatively low order L such that $L \ll n$; practical implementation indicates that typically $L \sim 3$ is often quite adequate [14–18]. If there is no cooperation between the input variables, then only zeroth order and first order terms will appear in the expansion. However, even to first order the expansion is not a linear superposition, as $f_i(x_i)$ could have an arbitrary dependence on x_i. The high dimensional approximation formulation in this paper is based on HDMR and an efficient means is presented for obtaining its component functions when the sample points are random. With scattered inputs, the determination of the component functions f_0, $\{f_i(x_i)\}$, $\{f_{ij}(x_i, x_j)\}$, etc., requires the evaluation of integrals involving $f(x)$, and they will be carried out with Monte Carlo integration since it is the most viable algorithm to carry out integrals in high dimensions.

If it is possible to perform controlled experiments with the system (i.e., if there is freedom over choosing the input values), then a so called cut-HDMR can be formulated to represent $f(x)$ in the form (2) in a computationally efficient way [13]. The computational cost refers to the number of experiments needed to construct the representation. The cost of the arithmetic operations is not included in this count since most often it will constitute a negligible portion of the overall cost; the common problem is the scarcity of the data. The cut-HDMR representation of $f(x)$ can be converted to the ANOVA decomposition of $f(x)$ without the need for additional experiments [13,18]. However, in some applications one can not control the experiments (i.e., the nature of the inputs are inherently random or they are arbitrarily scattered as in chaotic time-series data). This paper will deal with this latter situation where we will present a random sampled HDMR (referred to as RS-HDMR) algorithm which will give the ANOVA decomposition of a multivariate output with random or quasi-random data. The HDMR representations can also be used for knowledge discovery in large databases to understand the structures hidden in the database. In this fashion HDMR can be employed to sift the important variables contributing to a specific output. Furthermore, the HDMR representation can be used for predictive modelling to build a black-box representation relating a set of inputs to an output variable.

The paper is organized as follows. Section 2 presents the algorithm for computing RS-HDMR of $f(x)$. It will be shown that the RS-HDMR component functions can be constructed by the minimization of a weighted least-squares functional. In section 3, two applications will be used to illustrate the technique in the text. Applications of the HDMR concept to other fields by different sampling means is presented in [14–17]. In the context of sensitivity analysis of the variance, it is important to ascertain the contribution of each of the input variables, or groups of inputs, to the overall statistical variance of the output f. In this manner we will show how the algorithm presented in this paper can be used to compute the statistical sensitivity indices [19–22] in the analysis of

variance, and the efficiency of the algorithm will be shown to significantly exceed that of prior means. Although one may perform this uncertainty analysis without directly constructing a model, an HDMR derived model is fundamental to overall system analysis and it can be used for additional purposes. Along these lines, the second illustration is an inverse problem in discrete dynamical systems. An HDMR map will be found by using discrete time-series data, and such maps can be used for at least short-term predictive purposes. Concluding remarks are presented in section 4.

2. The algorithm

A specific member of the HDMR family in equation (2) is used in statistics as the ANOVA decomposition [10–12] of a multivariate statistical output $f(x) \equiv f(x_1, x_2, \ldots, x_n)$ which depends on independently distributed random inputs $x_1, x_2, \ldots, x_n$. A basic conceptual distinction between the ANOVA decomposition and HDMR concerns the use of the representation in equation (2). ANOVA only exploits the representation as a means of obtaining the variance components of the output, while HDMR is after the expansion functions as an input–output map for various purposes. Without loss of generality, we will assume that $f(x)$ takes values on the unit hypercube $[0, 1]^n$. The RS-HDMR decomposition of the function of the function $f(x)$ in equation (2) can be obtained via the minimization of the functional $\mathcal{J}$ below:

$$\varepsilon_\ell(x) \equiv f(x) - \left[f_0 - \sum_i f_i(x_i) - \sum_{i<j} f_{ij}(x_i, x_j) - \cdots - \sum_{i_1<\cdots<i_\ell} f_{i_1\cdots i_\ell}(x_{i_1}, \ldots, x_{i_\ell}) \right],$$

$$\mathcal{J} = \int_{[0,1]^n} \left[\varepsilon_\ell(x)\right]^2 \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_n \tag{3}$$

subject to the "null property" of the component functions as constraints:

$$\int_{[0,1]} f_{i_1\ldots i_\ell}(x_{i_1}, \ldots, x_{i_\ell})\, \mathrm{d}x_k = 0 \quad \text{for } k = i_1, i_2, \ldots, i_\ell. \tag{4}$$

The solution of the above minimization problem uniquely gives the following component functions [13]:

$$\begin{aligned}
f_0 &\equiv \int_{[0,1]^n} f(x)\, \mathrm{d}x, \\
f_i(x_i) &\equiv \int_{[0,1]^{n-1}} f(x) \prod_{j\neq i} \mathrm{d}x_j - f_0, \\
f_{ij}(x_i, x_j) &\equiv \int_{[0,1]^{n-2}} f(x) \prod_{k\notin\{i,j\}} \mathrm{d}x_k - f_i(x_i) - f_j(x_j) - f_0, \\
&\vdots \\
f_{i_1\ldots i_\ell}(x_{i_1}, \ldots, x_{i_\ell}) &\equiv \int_{[0,1]^{n-\ell}} f(x) \prod_{k\notin\{i_1,\ldots,i_\ell\}} \mathrm{d}x_k - \sum_{j_1<\cdots<j_{\ell-1}\subset\{i_1 i_2 \ldots i_\ell\}} f_{j_1\ldots j_{\ell-1}}(x)
\end{aligned}$$

$$- \sum_{j_1<\cdots<j_{\ell-2}\subset\{i_1 i_2 \ldots i_\ell\}} f_{j_1\ldots j_{\ell-2}}(x) - \cdots - \sum_j f_j(x_j) - f_0. \quad (5)$$

The null property in equation (4) serves to assure that the functions are orthogonal,

$$\int_{K^n} f_{i_1,\ldots,i_s}(x_{i_1},\ldots,x_{i_s}) f_{j_1,\ldots,j_p}(x_{i_1},\ldots,x_{i_p})\,\mathrm{d}x_1\,\mathrm{d}x_2\ldots\mathrm{d}x_n = 0 \quad (6)$$

for at least one index differing in $\{i_1,\ldots,i_s\}$ and $\{j_1,\ldots,j_p\}$, and s may be the same as p. Due to the orthogonality of the individual functions, the overall statistical variance of f is equal to the sum of the variances of the individual random variables on the right hand side of the equation (2). Usually only the lowest order terms (i.e., up to $L \sim 2$) have significant contributions to the overall variance of f and computation of them is sufficient for an accurate approximation of the variance of f. The RS-HDMR can be used in a broader context as one may be interested in more than the variance or some finite moment of the output. The decomposition may be used for representing the overall input–output relationship of the physical system. As such, the RS-HDMR expansion forms a multivariate approximation/interpolation scheme as well as a means for ANOVA to analyze the relevant statistics of a random output.

In this paper we assume that data is arbitrarily scattered throughout the domain of the input variables. This distinction is made as in many applications one has no freedom in the sampling of the input variables or the inputs are simply not control variables. If one has control over the input variables then the computationally efficient cut-HDMR [13,18] or other sampling strategies like Latin hypercube sampling may be applied for constructing an HDMR equivalent to the RS-HDMR decomposition of the output. In order to generate a cut-HDMR, the output is sampled on a Cartesian mesh throughout the domain of the input variables and when the high order correlated effects of the input variables upon the output are negligible, the number of sample values needed varies polynomically with the dimensionality of the function. The premise underlying HDMR's is the ansatz that high order correlated effects of the inputs upon the output are weak for most well defined physical systems [13]. The approximation of the output from the scattered data then reduces to low-dimensional approximation of the component functions, a task computationally much reduced over approximation of the high dimensional output. The RS-HDMR decomposition of the output $f(x)$ can be attained once the integrals in equation (5) are evaluated. There are two issues that need to be addressed to make this effort practical. The first one concerns the lack of control over the input variables. The integrals in equation (5) may be carried out via Monte Carlo sampling, but for scattered data it is not possible to efficiently compute them except for the grand average f_0. To be more specific, f_0 is approximated by

$$f_0 \approx \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{X}^i\right), \quad (7)$$

where N is the sample size, $\mathbf{X}^i = (x_1^i, x_2^i, \ldots, x_n^i)$ is the ith sample input and $f(\mathbf{X}^i)$ is the ith sample output value. The Monte Carlo approximation of the first order jth component function at an arbitrary point $y \in [0, 1]$ requires the following sum be evaluated:

$$f_j(y) \approx \frac{1}{N_y} \sum_{i=1}^{N_y} f\left(\mathbf{X}_y^i\right), \tag{8}$$

where N_y is the sample size and

$$\mathbf{X}_y^i = \left(x_1^i, \ldots, x_{j-1}^i, y, x_{j+1}^i, \ldots, x_n^i\right). \tag{9}$$

This procedure requires sampling the output on a regular net to compute the Monte Carlo approximations of the integrals in equation (5). The second issue concerns the cost of computing the above integrals. The number of model evaluations needed for accurately computing the value of a component function at a single point is of the order $N \gg 10^3$. This straightforward approach to the above integrals is not efficient. The algorithm we describe below does not directly compute the above integrals to get the component functions of RS-HDMR decomposition. We start with the following definition.

Definition. Let $\{\phi_{ik}(x_i)\}_{k=1}^s$ be a family of linearly independent approximating bases for the univariate functions of the variable x_i on the unit interval $[0, 1]$. They can be chosen as polynomials, orthogonal bases, splines, etc. We assume that each function in this family has zero mean, i.e.,

$$\int_{[0,1]} \phi_{ik}(x_i)\, \mathrm{d}x_i = 0, \quad k = 1, 2, \ldots, s. \tag{10}$$

If a family $\{\widehat{\phi}_{ik}(x_i)\}_{k=1}^s$ does not satisfy this condition it can be redefined to do so

$$\phi_{ik}(x_i) \equiv \widehat{\phi}_{ik}(x_i) - \int_{[0,1]} \widehat{\phi}_{ik}(x_i)\, \mathrm{d}x_i. \tag{11}$$

This new family $\{\phi_{ik}(x_i)\}_{k=1}^s$ satisfy the conditions (10). The approximating subspace $\mathcal{V}_i$ is defined as the linear span of this family, which we denote by $\mathcal{V}_i \equiv \mathrm{Span}\{\phi_{i1}(x_i), \ldots, \phi_{is}(x_i)\}$. In a similar manner, if $\{\phi_{ijk}(x_i, x_j)\}_{k=1}^s$ is a linearly independent approximating family for bivariate functions of the variables x_i, x_j, then we assume that they satisfy the following conditions:

$$\begin{aligned} \int_{[0,1]} \phi_{ijk}(x_i, x_j)\, \mathrm{d}x_i &= 0, \quad k = 1, 2, \ldots, s, \\ \int_{[0,1]} \phi_{ijk}(x_i, x_j)\, \mathrm{d}x_j &= 0, \quad k = 1, 2, \ldots, s. \end{aligned} \tag{12}$$

If the family $\{\widehat{\phi}_{ijk}(x_i, x_j)\}_{k=1}^{s}$ does not satisfy these conditions we can redefine a new family $\{\phi_{ijk}(x_i, x_j)\}_{k=1}^{s}$ to do so

$$\phi_{ijk}(x_i, x_j) \equiv \widehat{\phi}_{ijk}(x_i, x_j) - \int_{[0,1]} \widehat{\phi}_{ijk}(x_i, x_j)\, \mathrm{d}x_i - \int_{[0,1]} \widehat{\phi}_{ijk}(x_i, x_j)\, \mathrm{d}x_j + \int_{[0,1]^2} \widehat{\phi}_{ijk}(x_i, x_j)\, \mathrm{d}x_i\, \mathrm{d}x_j, \quad k = 1, 2, \ldots, s. \tag{13}$$

We denote their linear span by $\mathcal{V}_{ij} \equiv \mathrm{Span}\{\phi_{ij1}(x_i, x_j), \ldots, \phi_{ijs}(x_i, x_j)\}$. This construction easily can be generalized to any dimension $\ell \leqslant n$. We assume that the approximating family $\{\phi_{i_1 i_2 \cdots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell})\}_{k=1}^{s}$ satisfy the following conditions:

$$\int_{[0,1]} \phi_{i_1 i_2 \cdots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell})\, \mathrm{d}x_{i_m} = 0, \quad k = 1, 2, \ldots, s,\ m = 1, 2, \ldots, \ell. \tag{14}$$

The criteria in equations (10), (12) and (14) are fully consistent with their utilization to represent the functions as the L.H.S. of equation (2) and the orthogonality criterion in equation (6). Again the approximating subspace $\mathcal{V}_{i_1 \ldots i_\ell}$ is defined as the linear span of the family $\{\phi_{i_1 i_2 \cdots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell})\}_{k=1}^{s}$. A natural candidate for $\mathcal{V}_{i_1 \ldots i_\ell}$ is the ℓ-fold tensor product of the subspace $\mathcal{V}_i$ for any index i, i.e., $\mathcal{V}_{i_1 \ldots i_\ell}$ consists of linear combinations of the form

$$\phi_{i_1 i_2 \cdots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell}) = \phi_{i_1 k_1}(x_{i_1}) \phi_{i_2 k_2}(x_{i_2}) \cdots \phi_{i_\ell k_\ell}(x_{i_\ell}), \tag{15}$$

where $k \equiv (k_1, k_2, \ldots, k_\ell)$. The number of basis functions is taken as s which will depend on the order ℓ. We lastly define $\mathcal{V}_0$ as the subspace of constants, i.e., $\mathcal{V}_0 = \mathbb{R}^1$.

The following lemma uses these definitions.

Lemma. The variational problem below has a unique solution

$$\min_u \int_{[0,1]^n} \big[f(x) - u\big]^2\, \mathrm{d}x_1 \ldots\, \mathrm{d}x_n, \quad u \in \mathcal{V}_0 \oplus \sum_i \mathcal{V}_i \oplus \sum_{i<j} \mathcal{V}_{ij} \oplus \cdots \oplus \sum_{i_1 < i_2 \cdots < i_\ell} \mathcal{V}_{i_1 \ldots i_\ell}. \tag{16}$$

Proof. We will prove the assertion for $\ell = 2$. Generalization to higher dimensions is immediate by induction. For $\ell = 2$ the minimization functional is explicitly written as

$$\mathcal{J} = \int_{[0,1]^n} \left[f(x) - c_0 - \sum_{i=1}^{n} \sum_{k=1}^{s} c_{ik} \phi_{ik}(x_i) - \sum_{i<j}^{n} \sum_{k=1}^{s} c_{ijk} \phi_{ijk}(x_i, x_j) \right]^2 \mathrm{d}x. \tag{17}$$

From comparison with equation (3) it is evident that the component functions of the HMDR are expanded in the special sets of functions $\{\phi_{i_1 i_2 \cdots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell})\}$. We will calculate the coefficients c_0, $\{c_{ij}\}$ and $\{c_{ijk}\}$ uniquely to prove the lemma.

Setting the first variation of $\mathcal{J}$ to zero we get the first order conditions necessary for a minimum

$$\frac{\partial \mathcal{J}}{\partial c_0} = -2\int_{[0,1]^n}\left[f(x) - c_0 - \sum_{i=1}^{n}\sum_{k=1}^{s} c_{ik}\phi_{ik}(x_i) - \sum_{i<j}^{n}\sum_{k=1}^{s} c_{ijk}\phi_{ijk}(x_i, x_j)\right]\mathrm{d}x = 0. \quad (18)$$

Using identities (10) and (12) gives

$$-2\int_{[0,1]^n}\big[f(x) - c_0\big]\mathrm{d}x = 0, \quad c_0 = \int_{[0,1]^n} f(x)\,\mathrm{d}x. \quad (19)$$

Now, consider the first order coefficients $\{c_{ik}\}$

$$\frac{\partial \mathcal{J}}{\partial c_{pq}} = -2\int_{[0,1]^n}\left[f(x) - c_0 - \sum_{i=1}^{n}\sum_{k=1}^{s} c_{ik}\phi_{ik}(x_i) - \sum_{i<j}\sum_{k=1}^{s} c_{ijk}\phi_{ijk}(x_i, x_j)\right]\phi_{pq}(x_p)\,\mathrm{d}x. \quad (20)$$

Using identities (10) and (12) gives

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial c_{pq}} &= -2\int_{[0,1]^n}\left[f(x) - \sum_{k=1}^{s} c_{pk}\phi_{pk}(x_p)\right]\phi_{pq}(x_p)\,\mathrm{d}x \\ &= \int_{[0,1]^n} f(x)\phi_{pq}(x_p)\,\mathrm{d}x - \sum_{k=1}^{s} c_{pk}\int_{[0,1]}\phi_{pk}(x_p)\phi_{pq}(x_p)\,\mathrm{d}x_p = 0. \quad (21)\end{aligned}$$

Similarly, in the second order coefficients $\{c_{ijk}\}$ we get

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial c_{opm}} = -2\int_{[0,1]^n}\Bigg[&f(x) - c_0 - \sum_{i=1}^{n}\sum_{k=1}^{s} c_{ik}\phi_{ik}(x_i) \\ &- \sum_{i<j}\sum_{k=1}^{s} c_{ijk}\phi_{ijk}(x_i, x_j)\Bigg]\phi_{opm}(x_o, x_p)\,\mathrm{d}x.\end{aligned}$$

Using identities (10) and (12) gives

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial c_{opm}} &= -2\int_{[0,1]^n}\left[f(x) - \sum_{k=1}^{s} c_{opk}\phi_{opk}(x_o, x_p)\right]\phi_{opm}(x_o, x_p)\,\mathrm{d}x \\ &= \int_{[0,1]^n} f(x)\phi_{opm}(x_0, x_p)\,\mathrm{d}x - \sum_{k=1}^{s} c_{opk}\int_{[0,1]^2}\phi_{opk}(x_o, x_p)\phi_{opm}(x_o, x_p)\,\mathrm{d}x_o\,\mathrm{d}x_p \\ &= 0.\end{aligned}$$

The above equations can be solved for the coefficients c_0, $\{c_{ik}\}$ and $\{c_{ijk}\}$ if the family of matrices $\{\mathbf{M}^i\}$, $\{\mathbf{M}^{ij}\}$ defined below are invertible

$$\mathbf{M}^i_{k\ell} = \int_{[0,1]}\phi_{ik}(x_i)\phi_{i\ell}(x_i)\,\mathrm{d}x_i,$$

$$\mathbf{M}^{ij}_{k\ell} = \int_{[0,1]^2} \phi_{ijk}(x_i, x_j)\phi_{ij\ell}(x_i, x_j)\,\mathrm{d}x_i\,\mathrm{d}x_j. \tag{22}$$

The linear independence of the basis functions assures that matrices $\{\mathbf{M}^i\}$ and $\{\mathbf{M}^{ij}\}$ are invertible. The inverses of these matrices can be precomputed and stored, so they do not siginificantly add to the computational complexity of generating an RS-HDMR. The size of these matrices is expected to be comfortably small, assuming that the parent function $f(x)$ is well behaved. The only operations requiring sample values of the output involves computing the inner product of the function $f(x_1, \ldots, x_n)$ and the basis functions. Since the input data is assumed to be arbitrarily scattered, we will use Monte Carlo integrations to compute those quantities. Note that instead of evaluating the integrals in equation (5) which requires a costly special sampling of the output, we may efficiently compute the integrals utilizing arbitrarily scattered data. The Monte Carlo approximation for the above integrals is given by the following sums:

$$\begin{aligned}
c_0 &= \int_{[0,1]^n} f(x)\,\mathrm{d}x \approx \frac{1}{N}\sum_{r=1}^{N} f\left(x_1^{(r)}, x_2^{(r)}, \ldots, x_n^{(r)}\right),\\
c_{ij} &= \int_{[0,1]^n} f(x)\phi_{ij}(x_i)\,\mathrm{d}x \approx \frac{1}{N}\sum_{r=1}^{N} f\left(x_1^{(r)}, x_2^{(r)}, \ldots, x_n^{(r)}\right)\phi_{ij}\left(x_i^{(r)}\right),\\
c_{ijk} &= \int_{[0,1]^n} f(x)\phi_{ijk}(x_i, x_j)\,\mathrm{d}x \approx \frac{1}{N}\sum_{r=1}^{N} f\left(x_1^{(r)}, x_2^{(r)}, \ldots, x_n^{(r)}\right)\phi_{ijk}\left(x_i^{(r)}, x_j^{(r)}\right).
\end{aligned} \tag{23}$$

The inputs must be independent random variables for the convergence behavior of the Monte Carlo sums to hold. The appendix considers an analog of the development above assuming that the system is *a priori* defined with a discrete set of samples possibly over some sub-domain of the full space $[0, 1]^n$.

The HDMR conjecture coupled with the Monte Carlo approximation for the above integrals gives a computationally efficient representation of the output $f(x_1, \ldots, x_n)$. The errors made in Monte Carlo approximation of the coefficients are independent of the dimension n and of the order $1/\sqrt{N}$, with N being the sample size.

3. Applications

The HDMR technique is a tool to enhance mathematical modelling of a physical system where the interest centers on the input–output relationships. There is a broad family of applications that may exploit the HDMR capabilities [13], and two applications are given below of the algorithm presented in section 2 and the appendix.

3.1. Sensitivity analysis

Sensitivity analysis of a model aims to determine the importance of an input or a group of input variables upon the output. Sensitivity analysis of the output variance is

defined as the analysis of the contribution of each input variable or variable-group uncertainty to the overall variance of the output. If the random inputs consist of independently distributed uniform random variables, then the component functions will be uncorrelated and the overall variance can be decomposed according to equation (2) as follows:

$$\sigma = \mathbf{E}(f - f_0)^2 = \sum_i \sigma_i + \sum_{i<j} \sigma_{ij} + \cdots + \sigma_{12\ldots n}, \tag{24}$$

where the individual variances are given by

$$\sigma_{i_1 i_2 \ldots i_\ell} = \int_{[0,1]^\ell} (f_{i_1 \ldots i_\ell})^2 \, \mathrm{d}x_{i_1} \ldots \mathrm{d}x_{i_\ell}. \tag{25}$$

Global sensitivity indices based on these variances are defined as [22]

$$S_{i_1 i_2 \ldots i_\ell} = \frac{\sigma_{i_1 i_2 \ldots i_\ell}}{\sigma}, \tag{26}$$

where $S_{i_1 i_2 \ldots i_\ell}$ is the fractional contribution of the input set $\{x_{i_1}, \ldots, x_{i_\ell}\}$ to the variance of the output. In [22], the Monte Carlo approximation of the integrals in equation (25) was suggested, which is relevant to the sampling strategy introduced here. Computation of $\sigma_{i_1 i_2 \ldots i_\ell}$ requires the following sum to be evaluated

$$\frac{1}{N} \sum_{j=1}^{N} f(\mathbf{u}_j, \mathbf{x}_j) f(\mathbf{v}_j, \mathbf{x}_j), \tag{27}$$

where the input vector $\mathbf{x}_j$ contains the variables $x_{i_1}, x_{i_2}, \ldots, x_{i_\ell}$ for the jth sample where N is the sample size. The vectors $\mathbf{u}_j$ and $\mathbf{v}_j$ represent the remainder of the variables. The total number of random numbers which must be generated to compute all individual variances $\sigma_{i_1 i_2 \ldots i_\ell}$ up to the ℓth order is $N \times 2n$ and this number is independent of the order ℓ. The jth row of the random-number matrix provides the inputs for the product $f(\mathbf{u}_j, \mathbf{x}_j) f(\mathbf{v}_j, \mathbf{x}_j)$. The first n columns of the ith row provides the input vectors $\mathbf{u}_j$ and $\mathbf{x}_j$ and the remaining n columns provides only the input vector $\mathbf{v}_j$. The number of model evaluations to compute all the sums up to the Lth order is given by

$$N \times \sum_{i=0}^{L} \frac{n!}{(n-i)!i!}, \tag{28}$$

where a fixed sample size of N is used. The random numbers can be generated by either crude Monte Carlo or other forms of stratified sampling such as like Latin hypercube sampling.

There are two issues which limit the usefulness of the above sampling strategy. First, this strategy assumes that we can evaluate the output at predetermined sample points. They are random, but we require that the function be evaluated at those points. This may not be the case for applications where the inputs are still random but the data is just given and one does not have control over evaluating the model at predetermined points. Second, the computational cost given in equation (28) is very prohibitive at high

dimensions and given the ansatz that the HDMR decomposition is a dimension reduction technique, this number is unnecessarily large.

The algorithm given in section 2 can deal with scattered data and is computationally much less severe than the crude Monte Carlo evaluation of the sensitivity indices desribed above. One sample of size N is enough to compute all the sensitivities above and the sample size N is determined by the accuracy provided by Monte Carlo simulations. We give the following example for illustration.

The model is an analytical expression of three variables used previously in [22]

$$f(x_1, x_2, x_3) = \sin \pi x_1 + 7(\sin \pi x_2)^2 + 0.1\pi^4 (x_3)^4 \sin \pi x_1, \tag{29}$$

where the joint probability density of the inputs are given by

$$\begin{aligned} p(x_1, x_2, x_3) &= \prod_{i=1}^{3} p_i(x_i), \\ p_i(x_i) &= \begin{cases} \frac{1}{2} & \text{for } -1 \leqslant x_i \leqslant 1, \\ 0 & \text{for } x_i < -1,\ x_i > 1. \end{cases} \end{aligned} \tag{30}$$

The goal is to compute the global sensitivity indices $S_{i_1 i_2 \ldots i_\ell}$, $\ell \leqslant 3$, defined in equation (26). In [22] the same example is presented and a sample size of 1024 is used for the evaluation of each of the integrals $\sigma_{i_1 i_2 \ldots i_\ell}$ requiring 1024×7 evaluations to compute all the indices up to second order. We applied RS-HDMR to the above model at second order and used Legendre functions as the approximating bases. We set the number of basis functions at $s = 10$. The sensitivity indices corresponding to the sample sizes $N = 1024$ are presented in table 1. The first column indicates the exact sensitivities and the others correspond to sensitivities computed with the respective sample sizes.

RS-HDMR can also be used for extrapolation purposes as well. We carried out another experiment to assess this. A sample size of 1024 was used for constructing the HDMR. Chebyshev polynomials were used as the interpolator bases. We fixed the number of basis functions as $s = 10$. As the error criterion we have computed the following quantity for an ensemble of 1000 uniformly distributed points in $[0, 1]^3$:

$$\varepsilon \equiv \left\{ i = 1, \ldots, 1000\text{:} \ \left| \frac{f^i(x) - f^i_{\mathrm{HDMR}}(x)}{f^i(x)} \right| \right\} \tag{31}$$

Table 1

Indices	Exact	$N = 1024$	$N = 512$
σ_1	0.3138	0.3142	0.2998
σ_2	0.4424	0.4413	0.3930
σ_3	0.0	0.0	0.0
σ_{12}	0.0	0.0	0.0
σ_{13}	0.2436	0.2445	0.2872
σ_{23}	0.0	0.0	0.0
σ_{123}	0.0	0.0	0.0

Table 2

Sample size	1024	512	256	128
Mean[$\widehat{\varepsilon}$]	0.0761	0.1214	0.1489	0.2446
Median[$\widehat{\varepsilon}$]	0.0505	0.0768	0.1173	0.1578
Std.dev.[$\widehat{\varepsilon}$]	0.1002	0.1489	0.1623	0.2318
Length[$\widehat{\varepsilon}$]	981	977	954	911

(i.e., $i = 1, \ldots, 1000$ and $f^i(x)$ is the ith sample). However, this criterion can be misleading at the points where the model output is close to zero. We redefined the error vector ε again such that large values are not permitted. Since the frequency of these large values is also important, they were recorded too, i.e., we consider the following redefined error vector $\widehat{\varepsilon}$:

$$\widehat{\varepsilon} \equiv \{\varepsilon_i \colon \varepsilon_i < 1\} \tag{32}$$

and its length Length[$\widehat{\varepsilon}$]. The mean, median and the standard deviation about the mean of the above vector $\widehat{\varepsilon}$ corresponding to different sample sizes are shown in table 2. The length of the vector $\widehat{\varepsilon}$ is also given since it is a measure of the domain where the relative error is smaller than one. As expected the quality of the HDMR approximation is better with increasing sample size.

3.2. An inverse problem in dynamical systems

The forward problem in dynamical systems is to find the time evolution for a given initial state x_0. For a discrete dynamical system

$$x_n = f(x_{n-1}), \quad n = 1, 2, \ldots, \tag{33}$$

the vector $\{f^n(x_0)\}$, where $f^n(x_0) = f(x_n)$, gives this time evolution with the initial state x_0. The map f here denotes the mathematical model and most often we have incomplete knowledge of this map. The inverse problem of dynamical systems [23–25] is to find an approximation $\widehat{f}$ which will be close to the true map f generating the data $x_0, f(x_0), f^2(x_0), \ldots, f^N(x_0)$. The approximate map $\widehat{f}$ can be used as a predictive model for the system and we will use the criterion in equation (31) as a measure of the discrepancy between f and $\widehat{f}$. We used RS-HDMR to construct $\widehat{f}$ for the data generated by the following discrete dynamical system

$$\begin{aligned} x_n &= 0.671 - 0.416x_{n-1} - 1.014x_{n-1}^2 + 1.738x_{n-1}x_{n-2} \\ &\quad + 0.836x_{n-2} - 0.814x_{n-2}^2, \\ y_n &= x_n + \delta \end{aligned} \tag{34}$$

with the initial conditions

$$x_0 = 1.0, \qquad x_1 = 0.0. \tag{35}$$

Here y_n denotes the observed quantity and δ represents the measurement error or presence of noise in the system. We assume that the random variables δ's are independent,

Table 3

Std.dev.[δ]	0.05	0.1	0.25
Mean[$\widehat{\varepsilon}$]	0.1635	0.2227	0.3338
Median[$\widehat{\varepsilon}$]	0.0768	0.1324	0.2684
Std.dev.[$\widehat{\varepsilon}$]	0.2076	0.2315	0.2522
Length[$\widehat{\varepsilon}$]	464	430	368

identically distributed as a Gaussian with zero mean. An important issue here is to determine the minimal embedding dimension [26,27] of the data so that a dynamical system of the form in equation (33) can be written. We assume here that this dimension is already determined and dynamical system is of the form

$$x_n = f(x_{n-1}, x_{n-2}), \qquad y_n = x_n + \delta \tag{36}$$

consistent with equation (34) having the embedding dimension of two. This example serves to illustrate the inherently discrete algorithm in the appendix. An RS-HDMR was constructed to second order $L = 2$ by using Legendre functions as the approximating basis. We set the number of basis functions at $s = 10$. A sample size of $N = 512$ was used to construct $\widehat{f}$. The statistics for the error criterion in equation (31) is given in table 3 with different magnitudes of the noise term in the model. The quality of $\widehat{f}$ was evaluated with a test sample of $N = 512$ separate from the sample used to construct $\widehat{f}$. The error statistics is given for various values of the standard deviation of δ. As expected, the accuracy of the approximation decreases with increasing noise δ.

4. Conclusion

High dimensional model representations (HDMR) are a family of multivariate approximation techniques which are based on the ansatz that high order interaction effects of the input on the output is usually weak. Theoretical aspects of HDMR's are given in [13] and a computationally efficient HDMR is presented in [18]. This efficient HDMR required that output be sampled at specific points throughout the input domain. In most applications that is not the case. Inputs are distributed in a random or quasi-random fashion. The random-sample HDMR (RS-HDMR) presented in this paper handles this case in a computationally efficient way.

Two examples are chosen to illustrate the application of RS-HDMR. Sensitivity indices of a multivariate statistical quantity represent the relative contribution of each input(s) to the overall variance of the output. RS-HDMR was shown to be computationally very efficient to compute sensitivity indices with high accuracy. Predictive modelling for time series is an active area of research [26] for building a dynamical system from real-life data. RS-HDMR is a multivariate approximation scheme and as such it can be used to construct a data-generating dynamical system.

Acknowledgement

The authors acknowledge support from the Petroleum Research Fund of the American Chemical Society.

Appendix

In this appendix, we will give the HDMR solution to an inherently discrete formulation of the cost functional presented in section 2. We express the observed data as $f(\mathbf{x}^i) \equiv f(x_1^i, x_2^i, \ldots, x_n^i),\ i = 1, 2, \ldots, N$, where the index i denotes the ith sample and N stands for the sample size. By using the cost functional equation (17), the coefficients c_0, c_{ik}, c_{ijk}, etc. can be computed as integrals involving the parent function $f(x)$. Monte Carlo integration is the natural choice when the data is randomly scattered throughout the domain of $f(x)$. The randomness assumption allows for error estimates on the coefficients, and these error estimates will scale with $1/\sqrt{N}$, where N is the sample size. When the data is inherently non-random, a discrete cost functional of the form

$$\mathcal{J} = \sum_{m=1}^{N} \left[f\left(\mathbf{x}^m\right) - c_0 - \sum_{i=1}^{n} \sum_{k=1}^{s} c_{ik} \xi_{ik}\left(x_i^m\right) - \sum_{i<j} \sum_{k=1}^{s} c_{ijk} \xi_{ijk}\left(x_i^m, x_j^m\right) \right]^2 \tag{A.1}$$

can be used to compute the coefficients. However, the error estimates for these coefficients can not be readily estimated. Even if the HDMR converges at the Lth order, the coefficients computed by minimizing the functional in equation (A.1) gives an approximation which is good only where the data is clustered.

We start with the following definition.

Definition. Let $\{\widehat{\xi}_{ik}(x_i)\}_{k=1}^{s}$ be a family of approximating bases for the univariate functions of the variable x_i on the unit interval $[0, 1]$. These functions can be chosen as polynomials, orthogonal bases, splines, etc. We will redefine this family so that it obeys the following summability conditions

$$\xi_{ik}(x_i) \equiv \widehat{\xi}_{ik}(x_i) - \frac{1}{N} \sum_{\ell=1}^{N} \widehat{\xi}_{ik}\left(x_i^\ell\right). \tag{A.2}$$

This new family $\{\xi_{ik}(x_i)\}_{k=1}^{s}$ satisfy the following zero-sum condition:

$$\frac{1}{N} \sum_{\ell=1}^{N} \xi_{ik}\left(x_i^\ell\right) = 0. \tag{A.3}$$

The approximating subspace $\mathcal{V}_i$ is defined as the linear span of this family, which we denote by $\mathcal{V}_i \equiv \mathrm{Span}\{\xi_{i1}(x_i), \ldots, \xi_{is}(x_i)\}$. In a similar manner, if $\{\widehat{\xi}_{ijk}(x_i, x_j)\}_{k=1}^{s}$ is a

linearly independent approximating family for bivariate functions of the variables x_i, x_j, then we redefine them to satisfy the following conditions:

$$\xi_{ijk}(x_i, x_j) \equiv \widehat{\xi}_{ijk}(x_i, x_j) - \frac{1}{N}\sum_{\ell=1}^{N}\widehat{\xi}_{ijk}\big(x_i, x_j^{\ell}\big) - \frac{1}{N}\sum_{\ell=1}^{N}\widehat{\xi}_{ijk}\big(x_i^{\ell}, x_j\big) + \frac{1}{N^2}\sum_{\ell=1}^{N}\sum_{\ell'=1}^{N}\widehat{\xi}_{ijk}\big(x_i^{\ell}, x_j^{\ell'}\big). \tag{A.4}$$

This new family $\{\widehat{\xi}_{ijk}(x_i, x_j)\}_{k=1}^{s}$ satisfy the following zero-sum conditions:

$$\frac{1}{N}\sum_{\ell=1}^{N}\xi_{ijk}\big(x_i, x_j^{\ell}\big) = 0, \qquad \frac{1}{N}\sum_{\ell=1}^{N}\xi_{ijk}\big(x_i^{\ell}, x_j\big) = 0. \tag{A.5}$$

We denote their linear span by $\mathcal{V}_{ij} \equiv \mathrm{Span}\{\xi_{ij1}(x_i, x_j), \ldots, \xi_{ijs}(x_i, x_j)\}$. This construction can easily be generalized to any dimension $\ell \leqslant n$. We assume that the approximating family $\{\xi_{i_1 i_2 \cdots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell})\}_{k=1}^{s}$ satisfy the following conditions:

$$\frac{1}{N}\sum_{m=1}^{N}\xi_{i_1 i_2 \ldots i_\ell k}\big(x_{i_1}, \ldots, x_{i_p}^{m}, \ldots, x_{i_\ell}\big) = 0 \quad \text{for all } p,\ 1 \leqslant p \leqslant \ell. \tag{A.6}$$

Again the approximating subspace $\mathcal{V}_{i_1 \ldots i_\ell}$ is defined as the linear span of the family $\{\xi_{i_1 i_2 \cdots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell})\}_{k=1}^{s}$. A natural candidate for $\mathcal{V}_{i_1 \ldots i_\ell}$ is the ℓ-fold tensor product of the subspace $\mathcal{V}_i$ for any index i, i.e., $\mathcal{V}_{i_1 \ldots i_\ell}$ consists of linear combinations of the form

$$\xi_{i_1 i_2 \ldots i_\ell k}(x_{i_1}, \ldots, x_{i_\ell}) = \xi_{i_1 k_1}(x_{i_1})\xi_{i_2 k_2}(x_{i_2}) \cdots \xi_{i_\ell k_\ell}(x_{i_\ell}). \tag{A.7}$$

We lastly define $\mathcal{V}_0$ as the subspace of constants, i.e., $\mathcal{V}_0 = \mathbb{R}^1$.

The lemma below uses these definitions.

Lemma. The following variational problem has a unique solution.

$$\min_{u}\sum_{m=1}^{N}\big[f\big(\mathbf{x}^m\big) - u\big]^2, \quad u \in \mathcal{V}_0 \oplus \sum_{i}\mathcal{V}_i \oplus \sum_{i<j}\mathcal{V}_{ij} \oplus \cdots \oplus \sum_{i_1<i_2<\cdots<i_\ell}\mathcal{V}_{i_1 \ldots i_\ell}. \tag{A.8}$$

Proof. The proof is the discrete analogue of the proof given in section 2 and it will not be repeated here.

References

[1] I. Sobol, On the distribution of points in a cube and the approximate evaluation of each integrals, Comput. Math. Math. Phys. 7 (1976) 86–112.

[2] Y. Shreider, *The Monte Carlo Method* (Pergamon Press, Oxford, 1967).

[3] G.G. Lorentz, M.V. Golitschek and Y. Makovoz, *Constructive Approximation* (Springer, New York, 1996).
[4] F. Girosi and T. Poggio, Representation properties of networks: Kolmogorov's theorem is irrelevant, Neurocomputing 1 (1989) 465–469.
[5] J. Friedman and W. Stuetzle, Projection pursuit regression, J. Amer. Statist. Assoc. 76 (1981) 817–823.
[6] P. Diaconis and M. Shahshahani, On nonlinear functions of linear combinations, SIAM J. Sci. Statist. Comput. 5(1) (1984) 175–191.
[7] P. Huber, Projection pursuit, Ann. Statist. 13(2) (1981) 435–525.
[8] D. Parker, Learning logic, Working paper 47, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology (1985).
[9] T. Poggio and F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (1990) 1481–1497.
[10] Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[11] H. Scheffe, *The Analysis of Variance* (Wiley, New York, 1959).
[12] B. Efron and C. Stein, The jackknife estimate of variance, Ann. Statist. 9(3) (1981) 586–596.
[13] H. Rabitz and Ö.F. Alış, General foundations of high dimensional model representations, J. Math. Chem. 25 (1999) 197–233.
[14] J.A. Shorter, P.C. Ip and H. Rabitz, An efficient chemical kinetics solver using high dimensional model representation, J. Phys. Chem. A 103 (1999) 7192–7198.
[15] J.A. Shorter, P.C. Ip and H. Rabitz, Radiation transport simulation by means of fully equivalent operational model, Geophys. Res. Lett. (2000) in press.
[16] H. Rabitz and K. Shim, Multicomponent semiconductor material discovery guided by a generalized correlated function expansion, J. Chem. Phys. 111 (1999) 10640–10651.
[17] K. Shim and H. Rabitz, Independent and correlated composition behavior of material properties: application to energy band gaps for the $Ga_\alpha In_{1-\alpha} P_\beta As_{1-\beta}$ and $Ga_\alpha In_{1-\alpha} P_\beta Sb_\gamma As_{1-\beta-\gamma}$ alloys, Phys. Rev. B 58 (1998) 1940–1946.
[18] H. Rabitz, Ö.F. Alış, J. Shorter and K. Shim, Efficient input–output model representations, Comput. Phys. Comm. 117 (1999) 11–20.
[19] I. Sobol, Sensitivity estimates for nonlinear mathematical models, Math. Mod. Comp. Exp. 1 (1993) 407–414.
[20] A. Saltelli and I. Sobol, About the use of rank transformation in sensitivity analysis of model output, Reliab. Eng. Sys. Safety 50 (1995) 225–239.
[21] A. Saltelli and J. Hjorth, Uncertainty and sensitivity analyses of OH-initiated dimethylsulphide (DMS) oxidation kinetics, J. Atm. Chem. 21 (1995) 187–221.
[22] T. Homma and A. Saltelli, Importance measures in global sensitivity analysis of nonlinear models, Reliab. Eng. Sys. Safety 52 (1996) 1–17.
[23] M. Casdagli, Nonlinear prediction of chaotic time series, Phys. D 51 (1989) 52–98.
[24] J.P. Crutchfield and B.S. McNamara, Equations of motion from a data series, Complex Systems 1 (1987) 417–452.
[25] J.D. Farmer and J.J. Sidorowich, Predicting chaotic time series, Phys. Rev. Lett. 59 (1987) 845–848.
[26] J.Y. Campbell, A.W. Lo and A.C. Mackinlay, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, NJ, 1997) chapter 12.
[27] P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, Phys. D 9 (1983) 189–208.