J Math Chem (2013) 51:2654–2667
DOI 10.1007/s10910-013-0228-6

ORIGINAL PAPER

A novel method for multivariate data modelling: Piecewise Generalized EMPR

M. Alper Tunga · Metin Demiralp

Received: 12 June 2013 / Accepted: 10 July 2013 / Published online: 18 July 2013

Abstract A multivariate data modelling problem consists of a number of nodes with associated function values. Increase in multivariance urges us to use divide-and-conquer algorithms in modelling process of these problems. High dimensional model representation based methods can partition a given multivariate data set into less-variate data sets and have the ability of building a model through these partitioned data sets. Generalized HDMR (GHDMR) is one of these methods and it is known that it works well for dominantly and purely additive natures. Piecewise Generalized HDMR is an alternative method and was developed to increase the efficiency of GHDMR but the performance of the method for modelling multiplicative natures is still not sufficient and acceptable. This work aims to develop a new piecewise method based on enhanced multivariance product representation which works well for representing multiplicative natures.

Keywords High dimensional model representation · Multivariate data modelling · Interpolation · Multidimensional problems · Approximation

1 Introduction

Dealing with multivariate data including a number of nodes and the associated function values is an important issue in many research areas of basic sciences and engineering.

M. A. Tunga (✉)
Software Engineering Department, Faculty of Engineering, Bahçeşehir University,
Beşiktaş, 34349 Istanbul, Turkey
e-mail: alper.tunga@bahcesehir.edu.tr

M. Demiralp
Computational Science and Engineering Program, Informatics Institute,
İstanbul Technical University, Maslak, 34469 Istanbul, Turkey
e-mail: metin.demiralp@gmail.com

As the number of independent variables and the number of nodes increase, it becomes hard to determine an analytical structure for the problem because of mathematical and computational incapabilities. To this end, high dimensional model representation (HDMR) philosophy can be used as a divide-and-conquer method to bypass these disadvantages and to construct a new representation for the problem under consideration [1,2]. The method has a finite expansion composed of mutually orthogonal less-variate components of the given multivariate function and aims to uniquely determine the general structure of each component under a product type weight and a number of vanishing conditions defined through multiple integrations [1,2]. This philosophy is used to partition a given multivariate data set into less-variate data sets such as univariate, bivariate or higher variate data sets. However, the product type weight need in HDMR brings a restriction in multivariate data partitioning such as the given data set should have an orthogonal geometry. That is, the function values at all possible nodes of the problem domain should be known for the modelling process [3]. On the other hand, it is obvious that we cannot know all function values in most cases. This results in the development of another HDMR based method for partitioning process of multivariate data sets in which the function values are known only at arbitrarily distributed nodes of the problem domain.

For this purpose, Generalized HDMR (GHDMR) method was developed under a general type weight [4]. This method uses the standard HDMR expansion. First, the HDMR components of the general weight are obtained and then the GHDMR components of the given problem are determined. The HDMR and GHDMR methods use the same finite expansion which has an additive nature. Hence, as it is expected, the numerical implementations show that both methods work well in modelling dominantly and purely additive natures while the performance of the method becomes insufficient for the multiplicative natures [3,4]. Piecewise GHDMR is another method that makes this philosophy a better working algorithm for dominantly and purely multiplicative structures [5]. This method splits the domain into subdomains and builds an analytical structure for each subdomain which correponds to a piecewise system. However, it is still needed to develop a new method to get better representations for modelling multiplicative natures.

This work aims to take enhanced multivariance product representation (EMPR) into consideration for our modelling process [6]. This method provides a better analytical structure as the representation of a multivariate data modelling problem having multiplicative nature by inserting the GHDMR features into its algorithm [7]. The main purpose of this study is to improve the performance of the Generalized EMPR (GEMPR) method [7]. To this end, the proposed method is Piecewise GEMPR which splits the problem domain into subdomains, applies the steps of the new algorithm to each subdomain and constructs a piecewise structure by taking the representations obtained in each subdomain into consideration.

The HDMR philosophy is also used in modelling various scientific and engineering problems by many other scientists. Some of these research areas are about reliability analysis [8], helicopter aeroelastic analysis [9], laminar burning velocity [10], general formulation of HDMR component functions [11], random sampling [12], weight optimization [13], sensitivity analysis [14,15] and decision making [16].

The paper is organized as follows. The Sect. 2 gives the multivariate data descriptions that will help to define the data modelling problem. The Sect. 3 is about the mathematical background of the HDMR philosophy. The details of the GEMPR method are given in the Sect. 4 while the Sect. 5 covers Piecewise Generalized EMPR which is the proposed method of this work. The numerical implementations to test the performance of our new method are included in the Sect. 6. The Sect. 7 discusses the concluding remarks of the work.

2 Problem definition

This work aims to determine a structure as the analytical model of a given multivariate data. This multivariate data can be defined as

$$d_k \equiv \left(v_1^{(k)}, \ldots, v_N^{(k)}, \varphi_k\right), \quad \varphi_k \equiv f\left(v_1^{(k)}, \ldots, v_N^{(k)}\right), \quad 1 \le k \le m \tag{1}$$

where N and m stand for the number of independent variables and the number of the nodes that describe the given problem, respectively. This data set is also called "training data set" from which our method learn the problem and construct an analytical structure.

It is assumed that a testing data set is also given in the problem to examine the performance of the proposed method of this work. The general definition of this set can be written as

$$\mathscr{T}_\ell \equiv \left(\eta_1^{(\ell)}, \ldots, \eta_N^{(\ell)}\right), \quad 1 \le \ell \le t \tag{2}$$

where t is the total number of testing nodes. The original function value of each testing node will be compared with the function value obtained through our new method as the performance evaluation process. For this purpose, relative error formula is used in numerical implementations.

3 Mathematical background

The HDMR method has a finite expansion to express a given multivariate function, $f(x_1, \ldots, x_N)$, in terms of some functions having less number of independent variables. This expansion can be given as

$$f(x_1, \ldots, x_N) = f_0 + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) + \sum_{\substack{i_1, i_2 = 1 \\ i_1 < i_2}}^{N} f_{i_1 i_2}(x_{i_1}, x_{i_2}) + \cdots + f_{1\ldots N}(x_1, \ldots, x_N) \tag{3}$$

where N stands for the number of independent variables [1]. The most important task in the HDMR philosophy is to uniquely obtain the right hand side components of the above expansion under a product type weight with predefined orthogonality

conditions [1,2]. The product type weight prerequisite of HDMR restricts us to partition a multivariate data set into less variate data sets, that is, the fuction values at all possible nodes of the problem domain are needed to be known. However, the real life problems have multivariate training data set whose nodes are only a small subset of the whole problem domain. This makes HDMR unemployable on these types of data modelling problems. The GHDMR method uses a general weight function instead of a product type weight to bypass the mentioned disadvantage of HDMR [4]. To this end, the first step is to write the HDMR expansion of the general weight as

$$W(x_1,\ldots,x_N)=W_0+\sum_{i_1=1}^{N}W_{i_1}(x_{i_1})+\sum_{\substack{i_1,i_2=1\\ i_1<i_2}}^{N}W_{i_1i_2}(x_{i_1},x_{i_2})+\cdots+W_{1\ldots N}(x_1,\ldots,x_N) \tag{4}$$

To find the general structure of each HDMR component of the general weight function, the following auxiliary product type weight is defined with normalization criteria

$$\Omega(x_1,\ldots,x_N)\equiv\prod_{j=1}^{N}\Omega_j(x_j),\quad \int_{a_j}^{b_j}dx_j\,\Omega(x_j)=1 \tag{5}$$

A normalization criterion is also defined for the general weight since the HDMR philosophy requires normalized weights in its algorithm [4]

$$\int_{a_1}^{b_1}dx_1\cdots\int_{a_N}^{b_N}dx_N\left(\prod_{j=1}^{N}\Omega_j(x_j)\right)W(x_1,\ldots,x_N)=1 \tag{6}$$

To determine the HDMR components of the general weight function, the following orthogonality conditions are defined [4]

$$\int_{a_{i_\ell}}^{b_{i_\ell}}dx_{i_\ell}\Omega_{i_\ell}(x_{i_\ell})W_{i_1\ldots i_k}(x_{i_1},\ldots,x_{i_k})=0,\quad 1\le k\le N,\ 1\le\ell\le k,\ 1\le i_1<\cdots<i_k\le N \tag{7}$$

The general structure of the constant component, W_0 is obtained by using the following operator

$$\mathscr{I}_0F(x_1,\ldots,x_N)\equiv\int_{a_1}^{b_1}dx_1\Omega_1(x_1)\cdots\int_{a_N}^{b_N}dx_N\Omega_N(x_N)F(x_1,\ldots,x_N) \tag{8}$$

where $F(x_1, \ldots, x_N)$ is an arbitrary square integrable multivariate function. When the above operator is applied to the both sides of (4) under the normalization and orthogonality conditions given in (6) and (7), the following result for W_0 is obtained [4]

$$W_0 = 1 \tag{9}$$

Since the proposed method needs only the constant HDMR component of the general weight function, the structure of higher variate components are not included in this section.

4 The Generalized EMPR method

The EMPR method has the following finite expansion

$$f(x_1, \ldots, x_N) = f_0 \prod_{j=1}^{N} s_j(x_j) + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) \prod_{\substack{j=1 \\ j \neq i_1}}^{N} s_j(x_j) + \sum_{\substack{i_1, i_2 = 1 \\ i_1 < i_2}}^{N} f_{i_1 i_2}(x_{i_1}, x_{i_2}) \times \prod_{\substack{j=1 \\ j \neq i_1, i_2}}^{N} s_j(x_j) + \cdots + f_{1 \ldots N}(x_1, \ldots, x_N) \tag{10}$$

where each $s_j(x_j)$ function is called "support function" and N is the number of independent variables of the given multivariate function [6]. The method aims to construct a representation for the given analytical structure. In this sense, the selection of the support functions and the determination of the structure of the EMPR components appearing at the right hand side of the EMPR expansion are the main parts of the proposed algorithm of this work. This section covers the determination of the GEMPR components under the implicit form of the support functions. The selection process of these support functions is given in the next section.

To determine the general structure of each GEMPR component, normalization criteria on support functions and vanishing conditions under the general weight, auxiliary weight and the support functions for the mentioned components are defined as the first two steps of the algorithm [7]. The normalization criteria under the product type weight in EMPR method is defined as [6]

$$\int_{a_i}^{b_i} dx_i \, \Omega_i(x_i) s_i(x_i)^2 = 1, \quad 1 \leq i \leq N \tag{11}$$

The normalization criteria on support functions under the general weight can be obtained through the following N-tuple integration

$$\int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N \, W(x_1, \ldots, x_N) \Omega_i(x_i) s_i(x_i)^2 = 1, \quad 1 \leq i \leq N \tag{12}$$

while the vanishing conditions that help us to determine the GEMPR components of the expansion given in (10) are defined as

$$\int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N \left(\prod_{j=1}^{N} \Omega_j(x_j) \right) \left(\prod_{j=1}^{N} s_j(x_j)^2 \right) W(x_1, \ldots, x_N) f_i(x_i) = 0 \quad (13)$$

where $1 \leq i \leq N$.

The $\mathscr{I}_0$ operator given in (8) can be rewritten for the GEMPR case as

$$\mathscr{I}_0 F(x_1, \ldots, x_N) \equiv \int_{a_1}^{b_1} dx_1 \Omega_1(x_1) s_1(x_1) \cdots \int_{a_N}^{b_N} dx_N \Omega_N(x_N) s_N(x_N) F(x_1, \ldots, x_N) \quad (14)$$

This operator is used as follows to determine the constant GEMPR component of the given multivariate function

$$\mathscr{I}_0 [W(x_1, \ldots, x_N) f(x_1, \ldots, x_N)] = \mathscr{I}_0 \left[\left(W_0 + \sum_{i_1=1}^{N} W_{i_1}(x_{i_1}) + \cdots \right) \times \left(f_0 \prod_{j=1}^{N} s_j(x_j) + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) \prod_{\substack{j=1 \\ j \neq i_1}}^{N} s_j(x_j) + \cdots \right) \right] \quad (15)$$

When the relations (11), (12), (13) and (14) are taken into consideration, the constant component is obtained as

$$f_0 = \int_{a_1}^{b_1} dx_1 \Omega_1(x_1) s_1(x_1) \cdots \int_{a_N}^{b_N} dx_N \Omega_N(x_N) s_N(x_N) W(x_1, \ldots, x_N) f(x_1, \ldots, x_N) \quad (16)$$

Since it is known that the univariate GHDMR components are the unknowns of a system of linear equations which has sometimes no solutions because of linearly dependent equations [4] and the GEMPR method has a similar philosophy with GHDMR [7], this work aims to use only the constant component to bypass this disadvantage and to reduce the mathematical and computational complexity.

In addition, the main purpose here is to construct a model for a given multivariate data. In this sense, we need to define a general weight which has the ability of taking each training node with its associated function value into consideration in the modelling process [4]. To this end, the following Dirac delta type weight is selected as the general weight function

$$W(x_1,\ldots,x_N) \equiv \sum_{k=1}^{m} \alpha_k \delta\left(x_1 - \nu_1^{(k)}\right) \cdots \delta\left(x_N - \nu_N^{(k)}\right) \tag{17}$$

where α_k parameters are used for giving different importance to each individual datum. Relation (6) gives the following constraint on these α_k parameters

$$\sum_{k=1}^{m} \alpha_k \overline{\Omega}_k = 1, \quad \overline{\Omega}_k \equiv \prod_{j=1}^{N} \Omega_j\left(\nu_j^{(k)}\right), \quad 1 \le k \le m \tag{18}$$

The relation given in (16) can be reorganized by inserting the weight function given in (17) and the constant GEMPR component for the data partitioning procedure is obtained as

$$f_0 = \sum_{k=1}^{m} \alpha_k \overline{\Omega}_k \overline{s}_k \varphi_k, \quad \overline{s}_k \equiv \prod_{j=1}^{N} s_j\left(\nu_j^{(k)}\right), \quad 1 \le k \le m \tag{19}$$

To this end, the following constant GEMPR approximant is obtained as the approximation to the original function under consideration

$$f(x_1,\ldots,x_N) \approx \pi_0(x_1,\ldots,x_N) = f_0 \prod_{j=1}^{N} s_j(x_j) \tag{20}$$

The other important case is to select appropriate support functions to get a better approximation through the constant approximant for the given multivariate data modelling problem. There are no rules or algorithms to obtain the best support function structures for a given problem in literature. Besides, there is an experimental work about the influences of these support functions on the representation of a given analytical structure [6]. Our work is the first study in developing a new procedure to identify support functions that improves the quality of the constant EMPR approximant. This procedure is given in the next section which also covers the proposed method of this paper.

5 The Piecewise Generalized EMPR method

This work proposes a new algorithm based on GEMPR method which splits the whole domain of the given problem into subdomains and obtains an analytical structure in each subdomain. This new algorithm is called "Piecewise Generalized EMPR". The steps of this algorithm can be itemized as follows:

- Specify the total number of independent variables, N, of the given problem.
- Specify the total number of training nodes, m, of the given problem.
- Specify the domain of each independent variable as

$$x_i \in [a_i, b_i], \quad 1 \le i \le N \tag{21}$$

where a and b values stand for the minimum and the maximum values that the related independent variable can take.

- Find the number of different values that each independent variable can take in the problem domain and assign to $n_1, n_2, \ldots, n_N$ to be used in the support function determination process.
- Identify the total number of subintervals selected for the independent variables as $z_1, z_2, \ldots, z_N$.
- Determine the subintervals for each independent variable.

$$x_i^{(1)} \in \left[c_i^{(1)}, c_i^{(2)}\right), \; x_i^{(2)} \in \left[c_i^{(2)}, c_i^{(3)}\right), \; \cdots, \; x_i^{(z_i)} \in \left[c_i^{(z_i)}, c_i^{(z_i+1)}\right],$$
$$c_i^{(1)} \equiv a_i, \quad c_i^{(z_i+1)} \equiv b_i, \quad 1 \leq i \leq N \tag{22}$$

The interval of each independent variable is splitted into equal subintervals.

- Construct the subdomains of the whole problem domain through the cartesian product of the subintervals of the independent variables

$$\mathscr{D}^{(\rho)} \equiv x_1^{(j_1)} \times x_2^{(j_2)} \times \cdots \times x_N^{(j_N)},$$
$$1 \leq \rho \leq \zeta \quad 1 \leq j_1 \leq z_1, \; \cdots, \; 1 \leq j_N \leq z_N, \quad \zeta \equiv z_1 \times \cdots \times z_N \tag{23}$$

where ζ is the total number of the subdomains.

- Build the training data sets of each subdomain by taking the subintervals of each independent variable into consideration.
- Define an auxiliary weight function. In this work, the auxiliary weight is selected same as given in the study about GHDMR method [4]

$$\Omega(x_1, \ldots, x_N) = \prod_{j=1}^{N} \frac{1}{b_j - a_j} \tag{24}$$

Since the structure of this auxiliary weight depends on the minimum and maximum values that are included in the domain that is under consideration, the weight should be rewritten as follows for each subdomain

$$\Omega(x_1, \ldots, x_N)^{(\rho)} = \prod_{j=1}^{N} \frac{1}{\omega_j^{(\rho)} - \theta_j^{(\rho)}}, \quad 1 \leq \rho \leq \zeta \tag{25}$$

where $\omega_j^{(\rho)}$ and $\theta_j^{(\rho)}$ are corresponding to the upper and lower bounds of the related independent variable of the related subdomain, respectively [5].

- Select the support function structures. The most appropriate support function selection process is an important issue for EMPR based methods. When the analytical structure which is the model of the given problem is known, the support functions for that problem can be selected parallel to the factors of the given analytical structure [6]. However, this work aims to construct an analytical model of a problem in which some nodes of the problem domain with the associated function values are

given, that is, the analytical structure is asked to be determined. In this sense, one way to select the most powerful support functions that increases the performance of the EMPR based method is to develop an algorithm for the optimization process of support functions. This results in finding the solution of a nonlinear equations system which requires a complex mathematical and computational approach to get the results. Imposing a specific support function structure to the EMPR based method to model the given problem can be the second way of support function determination process. These structures may be polynomial, exponential, logarithmic and trigonometric. To this end, the following sets are specified for the support function families

$$\begin{aligned}
\mathscr{S}_1 &\equiv \left\{ s_j(x_j) = \left(1 + x_j\right)^{n_j - 1}, \quad 1 \le j \le N \right\} \\
\mathscr{S}_2 &\equiv \{ s_j(x_j) = \sin\left(x_j\right), \quad 1 \le j \le N \} \\
\mathscr{S}_3 &\equiv \{ s_j(x_j) = \cos\left(x_j\right), \quad 1 \le j \le N \} \\
\mathscr{S}_4 &\equiv \{ s_j(x_j) = \mathrm{e}^{\mathrm{x_j}}, \quad 1 \le \mathrm{j} \le \mathrm{N} \} \\
\mathscr{S}_5 &\equiv \{ s_j(x_j) = \ln\left(x_j\right), \quad 1 \le j \le N \} \\
\mathscr{S}_6 &\equiv \{ s_j(x_j) = x_j, \quad 1 \le j \le N \} \\
\mathscr{S}_7 &\equiv \{ s_j(x_j) = 1, \quad 1 \le j \le N \} \\
\mathscr{S}_8 &\equiv \{ s_j(x_j) = \beta^{x_j}, \quad \beta = 2, 3, \ldots, \quad 1 \le j \le N \}
\end{aligned} \tag{26}$$

where $n_1, n_2, \ldots, n_N$ are obtained in Step 4.

- Run the following steps for each support function family given above.
- Evaluate the constant GEMPR component and the related approximant for each subdomain by taking the relations (19) and (20) into consideration. To this end, a constant approximant is obtained in each subdomain as

$$\pi_0(x_1, \ldots, x_N)^{(\rho)} = f_0^{(\rho)} \prod_{j=1}^{N} s_j(x_j), \quad 1 \le \rho \le \zeta \tag{27}$$

where $f_0^{(1)}, f_0^{(2)}, \ldots, f_0^{(\zeta)}$ stand for the constant component of the first, second and the other subdomains, respectively since there may exist ζ number of subdomains in a given problem. Hence, we have a piecewise structure as the model of our problem. This can be represented by rewriting the relation (27) as

$$\pi_0(x_1, \ldots, x_N) = \begin{cases} f_0^{(1)} \prod_{j=1}^{N} s_j(x_j), & (x_1, x_2, \ldots, x_N) \in \mathscr{D}^{(1)} \\ f_0^{(2)} \prod_{j=1}^{N} s_j(x_j), & (x_1, x_2, \ldots, x_N) \in \mathscr{D}^{(2)} \\ \qquad\qquad \vdots & \\ f_0^{(\zeta)} \prod_{j=1}^{N} s_j(x_j), & (x_1, x_2, \ldots, x_N) \in \mathscr{D}^{(\zeta)} \end{cases} \tag{28}$$

- Evaluate the function value of each training node by using the constant Piecewise Generalized EMPR approximant obtained in the previous step as given in (28).

- Obtain the relative error value of each approximant determined by using different support function families given in (26).
- The support function will be selected by looking for the case in which the smallest relative error value obtained for the training process through the constant Piecewise Generalized EMPR approximant.
- Use the approximant having the smallest relative error in finding the function values of the testing nodes.

6 Numerical implementations

Several multivariate data modelling problems are constructed through a number of testing functions to examine the performance of the Piecewise Generalized EMPR method proposed in this work. In addition, these problems will allow us to compare the performance of our new method with GHDMR, Piecewise GHDMR and GEMPR performances. The evaluations are done by using MuPAD [17] within 20-digits precision. The data preparation process is executed through Perl [18] scripts. These Perl scripts organize the training data set to be easily used in MuPAD scripts.

The testing functions are selected as

$$
\begin{aligned}
&f_1(x_1,\ldots,x_5)=\prod_{i=1}^{5}x_i, \quad f_2(x_1,\ldots,x_5)=\left[\sum_{i=1}^{5}x_i\right]^{10},\\
&f_3(x_1,\ldots,x_5)=\left[\sum_{i=1}^{5}x_i\right]^{7}, \quad f_4(x_1,\ldots,x_5)=\left[\sum_{i=1}^{5}x_i\right]^{6},\\
&f_5(x_1,\ldots,x_5)=\left[\sum_{i=1}^{5}x_i\right]^{5}, \quad f_6(x_1,\ldots,x_5)=\sum_{i=1}^{5}x_i,\\
&f_7(x_1,\ldots,x_5)=\mathrm{e}^{\sum_{i=1}^{5}x_i}, \quad f_8(x_1,\ldots,x_5)=\sin\left(\prod_{i=1}^{5}x_i\right),\\
&f_9(x_1,\ldots,x_5)=\sin\left(\sum_{i=1}^{5}x_i\right), \quad f_{10}(x_1,\ldots,x_5)=\left[\cos\left(\sum_{i=1}^{5}x_i\right)\right]^{3}
\end{aligned}
\tag{29}
$$

where each has 5 independent variables. Each independent variable can take values from unit interval, [0, 1], for simplicity and generality while the independent variables, x_1, x_2, x_3, x_4 and x_5 take 5, 5, 4, 8, and 5 different values, respectively. This results in the following information needed to generate the first support function of the related family

$$
n_1=5, \quad n_2=5, \quad n_3=4, \quad n_4=8, \quad n_5=5 \tag{30}
$$

Table 1 Average relative error values for the training part

	GHDMR	PGHDMR	GEMPR	PGEMPR
$\mathcal{N}_{f_1}$	0.3082471450	0.1059598420	0.0647762246	**0.0516067839**
$\mathcal{N}_{f_2}$	0.4497975828	0.2007505725	0.3097779747	**0.1620777654**
$\mathcal{N}_{f_3}$	0.2684661626	0.1281011802	0.1522173606	**0.0916295342**
$\mathcal{N}_{f_4}$	0.2120456156	0.1046347004	0.1570680403	**0.1029420970**
$\mathcal{N}_{f_5}$	0.1514523942	**0.0818049184**	0.2068359074	0.1284423928
$\mathcal{N}_{f_6}$	**0.0**	0.0153817367	0.0305884685	0.0150725791
$\mathcal{N}_{f_7}$	0.0292018860	0.0288846951	0.0051703702	**0.0047151743**
$\mathcal{N}_{f_8}$	0.3090712351	0.1077620653	0.0568295004	**0.0515774525**
$\mathcal{N}_{f_9}$	0.0294573207	**0.0116139448**	0.0519885993	0.0278674899
$\mathcal{N}_{f_{10}}$	0.4859548271	0.2210190802	0.4224658408	**0.2197058837**

while the number of subintervals of each independent variable is as

$$z_1 = 2, \quad z_2 = 2, \quad z_3 = 2, \quad z_4 = 2, \quad z_5 = 2 \tag{31}$$

which gives totally 32 subdomains for the given problems. We must be careful about our decision on the number of these subintervals specified for each independent variable. If the total number of training nodes of given problem in a subdomain is very small, then this may cause a bad approximation for the corresponding subdomain.

We assume that we have 1,000 nodes with associated function values in each data modelling problem and we use 700 nodes of this data set in our training data set while 300 nodes are left for the testing data set. Both training and testing data sets are constructed randomly through a Perl script written by the authors. The performance of the proposed method of this study is examined by executing its algorithm for 30 randomly constructed multivariate data modelling problems through each testing function given in (29). The relative error for each execution of each problem is evaluated and then an average relative error value is obtained for each testing function.

When we run the Piecewise Generalized EMPR algorithm for each testing function given in (29), the support functions S_6, S_1, S_1, S_1, S_1, S_8 ($\beta = 2$), S_4, S_2, S_7 and S_1 are obtained for modelling the data sets constructed through the testing functions f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9 and f_{10}, respectively. It is observed that some of the support functions are not used for the given testing functions. This does not mean that we will not need these support functions in other implementations. One may remove those support functions from the list given in (26) or insert some new structures to have a more detailed list. When the best appropriate support function determination process finishes, the piecewise analytical structure is obtained as the model of the given testing function. This piecewise function can then be used to evaluate the function value at the given testing node of the multivariate data modelling problem under consideration.

Table 2 Average relative error values for the testing part

	GHDMR	PGHDMR	GEMPR	PGEMPR
$\mathcal{N}_{f_1}$	0.3095432620	0.1073527460	0.0657628930	**0.0560283091**
$\mathcal{N}_{f_2}$	0.4532413638	0.2128220358	0.3206331405	**0.1676329180**
$\mathcal{N}_{f_3}$	0.2781259189	0.1365285378	0.1580577570	**0.0945007450**
$\mathcal{N}_{f_4}$	0.2159593689	0.1099058725	0.1609143515	**0.1044185505**
$\mathcal{N}_{f_5}$	0.1585280015	**0.0856824499**	0.2091417497	0.1299247680
$\mathcal{N}_{f_6}$	**0.0**	0.0163393383	0.0311006072	0.0153576375
$\mathcal{N}_{f_7}$	0.0304306026	0.0304236412	0.0051703703	**0.0048594242**
$\mathcal{N}_{f_8}$	0.3148471414	0.1081448027	0.0587309701	**0.0553526190**
$\mathcal{N}_{f_9}$	0.0311129911	**0.0122045493**	0.0528750921	0.0291392325
$\mathcal{N}_{f_{10}}$	0.4987139635	0.2295387333	0.4448768481	**0.2210190802**

Table 1 shows the average relative error values for the training process of each testing function over 30 random runs. The error values in boldface stand for the best result obtained through the methods used in comparison. The results indicate that the Piecewise Generalized EMPR (PGEMPR) method which is the proposed method of this study gives the best approximation for functions having dominantly or purely multiplicative nature. These functions stand for the testing functions, f_1, f_2, f_3 and f_4. The performance of the method gets better while the multiplicativity dominance increases. In addition, the proposed method also works better for exponential functions, trigonometric functions whose argument is of type multiplicative nature and the powers of trigonometric functions. The testing functions, f_7, f_8 and f_{10} are used for these mentioned cases. On the other hand, the GHDMR method works well for the testing function, f_6 which has purely additive nature while Piecewise Generalized HDMR (PGHDMR) gets better approximations for dominantly additive natures and trigonometric functions that have arguments of type additive structure. The testing functions, f_5 and f_9, correspond to these two cases, respectively. The results obtained through the GEMPR method shows that we do not need to use that method for any case if we use the proposed method of this work. Our new method works better than GEMPR for all cases.

The discussions for the results of Table 1 are also true for the testing part as it is seen in Table 2. The best values are again boldface highlighted. The performance of the methods act same as we examine in the training part. This means that if the method is successful in learning the analytical model of the given multivariate data modelling problem from its training nodes then it can estimate the function value of the testing nodes of the problem by using that analytical model successfully.

Another important point is the stability of the proposed method which can be measured by evaluating the standard deviation value of each error value set obtained for each testing function. These values are given in Table 3. It can be examined that the results are very close to 0 which means the proposed method is very stable, that is, in each run through 30 randomly constructed problems the obtained relative error value is very closed to the others.

Table 3 Standard deviation values of the testing results

	GHDMR	PGHDMR	GEMPR	PGEMPR
$\mathcal{N}_{f_1}$	0.0196079741	0.0243787735	0.0484859645	0.0101366172
$\mathcal{N}_{f_2}$	0.0357442673	0.0390292083	0.0245028233	0.0220215646
$\mathcal{N}_{f_3}$	0.0204738391	0.0207787245	0.0192134006	0.0108541888
$\mathcal{N}_{f_4}$	0.0135314004	0.0125522265	0.0122305682	0.0090082536
$\mathcal{N}_{f_5}$	0.0115623714	0.0076963753	0.0132226753	0.0084918925
$\mathcal{N}_{f_6}$	0.0	0.0018559786	0.0023296255	0.0008396358
$\mathcal{N}_{f_7}$	0.0052065811	0.0026775012	0.0015710623	0.0007310763
$\mathcal{N}_{f_8}$	0.0218830030	0.0168249505	0.0349838160	0.0078625837
$\mathcal{N}_{f_9}$	0.0014702622	0.0010809233	0.0035254067	0.0017681133
$\mathcal{N}_{f_{10}}$	0.0175420259	0.0283769039	0.0251365072	0.0156985948

7 Concluding remarks

An analytical structure determination process for multivariate data modelling problems is an important concern in many research areas. One way for this purpose is to use GHDMR method which is a divide-and-conquer algorithm. The main philosophy of that method is to partition the given multivariate data set into univariate data sets and then interpolate these univariate data sets instead of interpolating a single multivariate data set. This reduces the complexity of the modelling process. Because GHDMR uses the HDMR expansion which is of type additive nature, the method works well for the multivariate data modelling problems having dominantly or purely additive nature. As the multiplicativity nature of the given problem becomes dominant, the performance of the method becomes poor. This work proposes a new method which is called Piecewise Generalized enhanced multivariance product representation (PGEMPR) to overcome this disadvantage.

Enhanced multivariance product representation (EMPR) is a recently developed method to represent multivariate functions in terms of less variate functions to reduce the mathematical and computational complexities. The support functions appearing in the expansion of EMPR let the method works well for especially functions having dominantly or purely multiplicative nature. Since EMPR is a multivariate function representation method, this technique can also be used for multivariate data modelling process. In this work, we develop a new EMPR based method which improves the performance of classical EMPR algorithm and has the ability of partitioning a given multivariate data into less variate data sets to obtain an approximate analytical structure as the model of the given data modelling problem. To construct this type of an algorithm we took the GHDMR philosophy into consideration and developed GEMPR method for data partitioning. The proposed method of this work is a piecewise based GEMPR which increases the performance level of GEMPR in data modelling. The determination process of the support functions affects the performance of the proposed method directly. The studies on optimizing these support functions show us that we should solve nonlinear equations system which is not preferred in numerical

implementations. To bypass this disadvantage we propose a more simple but working algorithm for obtaining appropriate support functions to model the given problem. The numerical results show us that this algorithm allows us to get acceptable approximations through Piecewise Generalized HDMR and to have better and stable results when compared with the other techniques appearing in the literature.

References

1. I.M. Sobol, Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exp. (MMCE) **1**, 407–414 (1993)
2. M. Demiralp, High dimensional model representation and its application varieties. Math. Res. **9**, 146–159 (2003)
3. M.A. Tunga, M. Demiralp, A new approach for data partitioning through high dimensional model representation. Int. J. Comput. Math. **85**, 1779–1792 (2008)
4. M.A. Tunga, M. Demiralp, Data partitioning via generalized high dimensional model representation (ghdmr) and multivariate interpolative applications. Math. Res. **9**, 447–462 (2003)
5. M.A. Tunga, M. Demiralp, Multivariate data modelling through piecewise generalized hdmr method. J. Math. Chem. **50**, 1711–1726 (2012)
6. B. Tunga, M. Demiralp, The influence of the support functions on the quality of enhanced multivariance product representation. J. Math. Chem. **48**, 827–840 (2010)
7. M.A. Tunga, M. Demiralp, Generalized enhanced multivariance product representation for data partitioning: constancy level. AIP Conf. Proc. **1389**, 1152–1155 (2011)
8. A.S. Balu, B.N. Rao, Inverse structural reliability analysis under mixed uncertainties using high dimensional model representation and fast fourier transform. Eng. Struct. **37**, 224–234 (2012)
9. S. Murugan, R. Chowdhury, S. Adhikari, M.I. Friswell, Helicopter aeroelastic analysis with spatially uncertain rotor blade properties. Aerosp. Sci. Technol. **16**, 29–39 (2012)
10. Z. Zhao, Z. Chen, Hdmr correlations for the laminar burning velocity of premixed $ch_4/h_2/o_2/n_2$ mixtures. Int. J. Hydrogen Energy **37**, 691–697 (2012)
11. G. Li, H. Rabitz, General formulation of hdmr component functions with independent and correlated variables. J. Math. Chem. **50**, 99–130 (2012)
12. G. Li, S.-W. Wang, H. Rabitz, Practical approaches to construct RS-HDMR component functions. J. Phys. Chem. A **106**, 8721–8733 (2002)
13. B. Tunga, M. Demiralp, Constancy maximization based weight optimization in high dimensional model representation for multivariate functions. J. Math. Chem. **49**, 1996–2012 (2011)
14. T. Ziehn, A.S. Tomlin, A global sensitivity study of sulfur chemistry in a premixed methane flame model using HDMR. Int. J. Chem. Kinet. **40**, 742–753 (2008)
15. T. Ziehn, A.S. Tomlin, Gui-hdmr—a software tool for global sensitivity analysis of complex models. Environ. Model. Softw. **24**, 775–785 (2009)
16. I. Banerjee, M.G. Ierapetritou, Model independent parametric decision making. Ann. Oper. Res. **132**, 135–155 (2004)
17. W. Oevel, F. Postel, S. Wehmeier, J. Gerhard, *The MuPAD Tutorial* (Springer, New York, 2000)
18. H.M. Deitel, P.J. Deitel, T.R. Nieto, D.C. McPhie, *How to Program Perl* (Prentice Hall, New Jersey, 2001)