Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 128 (2016) 42–54
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

Global sensitivity analysis using sparse high dimensional model representations generated by the group method of data handling

Romain S.C. Lambert[a,*], Frank Lemke[b], Sergei S. Kucherenko[a], Shufang Song[a,c], Nilay Shah[a]

[a] *Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK*
[b] *KnowledgeMiner Software, 13187 Berlin, Germany*
[c] *School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China*

Received 11 September 2014; received in revised form 14 January 2016; accepted 12 April 2016
Available online 2 May 2016

Highlights

- The group method of data handling (GMDH) is used to construct the high dimensional model representation (HDMR) to calculate Sobol's first and second order global sensitivity analysis indices.
- This methodology uses the parameter selection features of GMDH to construct a sparse HDMR expansion for high dimensional problems from a limited number of function evaluations.
- By design, the method also allows for the optimal (i.e. balancing accuracy and complexity) polynomial order selection in the HDMR expansion.

Abstract

In this paper, the parameter selection capabilities of the group method of data handling (GMDH) as an inductive self-organizing modelling method are used to construct sparse random sampling high dimensional model representations (RS-HDMR), from which the Sobol's first and second order global sensitivity indices can be derived. The proposed method is capable of dealing with high-dimensional problems without the prior use of a screening technique and can perform with a relatively limited number of function evaluations, even in the case of under-determined modelling problems. Four classical benchmark test functions are used for the evaluation of the proposed technique.

Keywords: Global sensitivity analysis; High dimensional model representations; Sobol indices; Group method of data handling

1. Introduction

Over the last decade, global sensitivity analysis (GSA) has gained considerable attention among practitioners, due to its advantages over local sensitivity analysis methods [4,8,9,21,20,29], for example in the detection of parameter

* Corresponding author. Tel.: +44 0 207 594 6611.
E-mail address: rl508@ic.ac.uk (R.S.C. Lambert).

http://dx.doi.org/10.1016/j.matcom.2016.04.005

interactions. An eminent class of GSA techniques is that of variance based methods, which includes the well-known Sobol method of global sensitivity indices [23]. Sobol sensitivity indices are used to rank input parameters and to discard unessential parameters. One way to reduce the computational expense of performing a sensitivity analysis is the use of surrogate-models or meta-models, which emulate the behaviour of the original computationally expensive models. Various surrogate modelling methods such as gaussian process modelling [16,18], polynomial chaos expansion (PCE) [2,3,25], and random sampling-high dimensional model representations (RS-HDMR) [14,15] have been proposed. RS-HDMR was originally defined as a set of quantitative tools to map the input–output behaviour of high dimensional systems [19]. Recently this technique has become popular and widely used by practitioners (e.g. [5,6,28,30–32]). RS-HDMR has also been used as an efficient way to compute first and second order Sobol global sensitivity indices. Despite improvements over the direct Sobol method, there have been attempts to create methods that can more efficiently generate (sparse) HDMR expansions. One particularly successful technique proposed by Blatman and Sudret [2,3] consists of the calculation of the polynomial chaos expansion (PCE) via a least angle regression (LAR) using cross validation schemes. Other adaptive methods to efficiently calculate HDMR expansions using machine learning techniques have also been suggested in the literature [7,28,30]. Ziehn and co-workers proposed an approach to calculate the 'optimal' polynomial order in HDMR expansions [30,31]. For completeness the reader is also invited to refer to bias correction methods [13] for the calculation of sensitivity indices [26].

In this paper we propose a new alternative method using techniques from the class of the so-called inductive modelling methods, namely the group method of data handling (GMDH). The group method of data handling (GMDH) was originally developed by Ivakhnenko and co-workers [10–12]. It is based on the principle of inductive self-organization. Unlike many other machine-learning techniques, this method is inductive which means that it does not a priori postulate the structure of the expressions. During model self-organization, GMDH generates, validates, and selects many alternative networks of growing complexity (i.e. with increasing number of parameters, interactions between these parameters and/or nonlinearity) until an 'optimally' complex model has been found (i.e. when it begins to over-fit the design data). A class of elementary expressions is used, which by making them gradually more complex, can describe every possible instance of a sought general function. GMDH has the ability to perform efficiently with a limited number of function evaluations in high dimensional spaces (under-determined systems), by selecting important parameters in an adaptive fashion (feature selection). A second key principle developed and introduced by the GMDH inductive modelling theory in the 1970s, and subsequently adopted for use in neural networks and other machine learning methods, is the principle of integrating external information into modelling to allow the objective selection of a model of optimal complexity [17,24].

We propose to use the characteristics of GMDH to efficiently calculate a sparse HDMR expansion, and to subsequently calculate Sobol first and second order global sensitivity indices.

This paper is organized as follows: the mathematical fundamentals of global sensitivity analysis and HDMR are introduced in Section 2, the GMDH method and the combined GMDH-HDMR method are presented in Section 3. The developed methodology is then applied to well-known benchmark functions in Section 4.

2. Methodology and statistical fundamentals

2.1. Sobol method of global sensitivity analysis

Consider an integrable function f defined in the unit hypercube $[0, 1]^M$. This function can be decomposed as

$$f(x) = \sum_{\alpha \subseteq \{1,\ldots,M\}} f_\alpha(x_\alpha). \tag{1}$$

Here α is a subset of indices from $\{1, \ldots, M\}$. A generic point of $[0, 1]^M$ is noted $x = (x_1, \ldots, x_M)$. $|\alpha|$ denotes the cardinality and x_α represents the $|\alpha|$-vector of components x_j, for $j \in \alpha$ Decomposition (1) is unique if

$$\forall i \epsilon \alpha, \quad \int_0^1 f_\alpha(x_\alpha)\, dx_i = 0, \tag{2}$$

in which case it is called the ANOVA decomposition. It follows from condition (2) that the ANOVA decomposition is orthogonal.

Furthermore we assume that f is square integrable over $[0, 1]^M$ Each component function $f_\alpha(x)$ is associated with a partial variance:

$$D_\alpha = \int f_\alpha(x)^2 dx. \tag{3}$$

Due to the orthogonality of the ANOVA decomposition the total variance D of function f is decomposed as:

$$D = \sum_{|\alpha|>0} D_\alpha. \tag{4}$$

Sobol's main effect sensitivity indices are defined as [18]:

$$S_\alpha = \frac{D_\alpha}{D}. \tag{5}$$

From (3)–(5) it follows that:

$$\sum_{|\alpha|>0} S_\alpha = 1. \tag{6}$$

Considering a set of parameters $\{1, \ldots, M\}$, a corresponding $\alpha \subseteq \{1, \ldots, M\}$, α' the subset of complementary parameters ($\{1, \ldots, M\} \setminus \alpha$), and using the previous definition of the variance we can compute the total variance of the subset α [23]:

$$D_\alpha^{tot} = D - D_{\alpha'}. \tag{7}$$

The total effect indices are defined as [18]:

$$S_\alpha^{tot} = \frac{D_\alpha^{tot}}{D}. \tag{8}$$

There are direct formulas for computing sensitivity indices using Monte-Carlo or Quasi Monte-Carlo integration (see e.g. [14]) but a more efficient approach for computation of sensitivity indices is based on building metamodels. In this paper we will focus on HDMR metamodels.

2.2. High dimensional model representation

Rabitz and co-workers [15,19] postulated that in many engineering problems the ANOVA decomposition of model functions (1) can be truncated to a sum of single effects and interactions of two parameters (or sometimes three parameters):

$$h(x) = \sum_{\alpha \subseteq \{1,\ldots,M\}, |\alpha| \leq 2} f_\alpha(x) = f_0 + \sum_{i=1}^{M} f_i(x_i) + \sum_{1 \leq i < j \leq M} f_{ij}(x_i, x_j). \tag{9}$$

Here h is an approximation of function f in (1). This decomposition is also known as ANOVA-HDMR.

The RS-HDMR method is based on the decomposition of low order component functions using local splines or orthogonal polynomials. Here we only consider the case of orthogonal polynomials $\{\varphi_p\}_{p \in \mathbb{N}^*}$. Low order component functions can be approximated as:

$$\begin{aligned} f_i(x_i) &\approx \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i), \\ f_{ij}(x_i, x_j) &\approx \sum_{p=1}^{l} \sum_{q=1}^{m} \beta_{pq}^{ij} \varphi_p(x_i) \varphi_q(x_j), \end{aligned} \tag{10}$$

where k, l, m represent predefined polynomial orders.

In this paper, we consider the shifted Legendre polynomials because the uncertain parameters are assumed uniformly distributed on the unit hypercube. With a sufficiently large number of samples N, decomposition coefficients in (9) can be computed by projecting the original function on the shifted Legendre polynomial basis:

$$\begin{aligned} &\forall r\epsilon\{1,\ldots,k\}, \quad \alpha_r^i = \int_0^1 f(x)\,\varphi_r(x_i)dx \approx \sum_{s=1}^{N} f(x)\,\varphi_r(x_i), \\ &\forall p\epsilon\{1,\ldots,l\},\ \forall q\epsilon\{1,\ldots,m\}, \quad \beta_{pq}^{ij} = \int_0^1 f(x)\,\varphi_p(x_i)\varphi_q\left(x_j\right)dx \\ &\qquad\qquad \approx \frac{1}{N}\sum_{s=1}^{N} f(x)\,\varphi_p(x_i)\varphi_q\left(x_j\right). \end{aligned} \tag{11}$$

Another way to calculate these coefficients is the use of regression [19]. Once decomposition coefficients are determined, it is easy to compute the Sobol sensitivity indices for single effects and interactions of two parameters using the coefficients from (11):

$$D_i = \sum_{r=1}^{k}\left(\alpha_r^i\right)^2, \tag{12}$$

$$D_{ij} = \sum_{p=1}^{l}\sum_{q=1}^{m}\left(\beta_{pq}^{ij}\right)^2. \tag{13}$$

The Sobol sensitivity indices can then be obtained by dividing (12)–(13) by the total variance of the output. There were many studies in which methods for efficient construction of HDMR expansions were suggested (see e.g. [28,30,31]). In the next section we show how inductive modelling methodologies and GMDH in particular, can offer an efficient way of calculating the RS-HDMR expansion.

3. Composing HDMR using GMDH

3.1. Multilayered iterative GMDH

In this section we give a brief outline of the working scheme of GMDH. For a more ample theoretical description of the method the reader is advised to refer to the work by Müller and co-workers [17]. We are particularly interested in an algorithm based on a *multi-layered* structure with iterative induction and selection steps introduced by Ivakhnenko [11]. An important property of this algorithm (especially in the case of the generation of an RS-HDMR model) is that it is very efficient for solving so-called under-determined modelling tasks where the number of samples is smaller than the number of potential inputs. The input data sample is a matrix containing N observations of a set of M parameters $x_i \ldots x_M$. The general procedure of GMDH consists of both a structure and a parameter estimation problem: Given input parameters x_i, output y, an initial model structure f consisting of the mean value of the output, an algorithm for inductively evolving model structure f, and an external selection criterion for evaluation and validation of usefulness of the evolved model structure f (Fig. 1), a first layer is then built by considering every possible parameter pair and inductively self-constructing neurons made of simple expressions on training data and validating them on testing data (external selection criterion or out-of-sample validation). These simple neuron expressions are typically within the frame of linear, multilinear functions or second order polynomials. For a multilinear model, which is the structure of GMDH neuron expression we are interested in this paper, for any pairs of inputs x_i x_k the first layer is of the following form:

$$\forall i,\ k\in\{1,\ldots,M\};\ i\neq k;\ \forall j\in\left\{1,2\ldots,\binom{M}{2}\right\}; \quad y_j^1 = \psi_j(x_i,x_k) = a_0 + a_1x_i + a_2x_k + a_3x_ix_k. \tag{14}$$

Here a_0, a_1, a_2 and a_3 are scalar model coefficients.

Each neuron function ψ_j will form part of a pool of competing candidate models consisting of a single neuron each for the first network layer (Fig. 2).

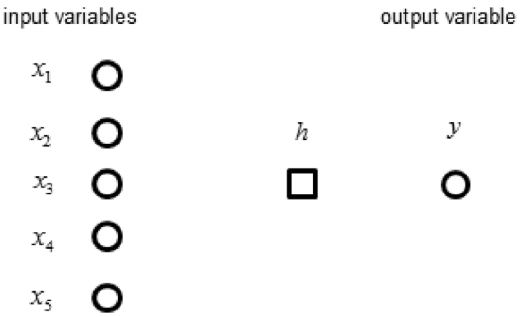

Fig. 1. Initial state of the multi-layered GMDH network model: model structure and parameters are undefined. For illustration purposes we consider a case of $\boldsymbol{M} = \mathbf{5}$.

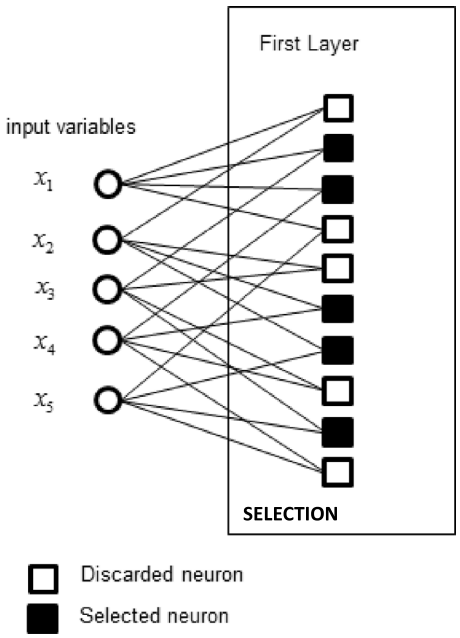

Fig. 2. First layer of competing neurons obtained with the multi-layered GMDH modelling process ($\boldsymbol{M} = \mathbf{5}$).

A number of best-fitting and best generalizing neurons are then selected according to the models' external selection criterion value i.e., the selection is performed based on the goodness of fit on a separate part of the data sample not used for training the neurons (Fig. 2). This external selection step avoids bias and over-fitting as model accuracy and its generalization power depend on model complexity and the structural uncertainty of the output,[1] and they become mutually exclusive properties beyond a critical point ('model of optimal complexity' [24]). This external model validation (hypothesis testing) is performed after each single induction step (hypothesis generation) as an integrated critical part of inductive model self-organization.

These selected intermediate models, in the classical approach, are subsequently used as inputs to create a new layer of more complex models while other models (neurons) are discarded during model selection (Fig. 3). The first layer validates models from information contained in any combination of two parameters of the dataset. The second layer uses information from up to four columns (initial parameters and their resulting combination in the first layer), the third from up to eight columns and so forth. In the kth layer, those m_{k-1} models selected in the preceding layer will

[1] In the case of RS-HDMR the structural uncertainty would be artificially created by excluding interactions of more than two parameters.

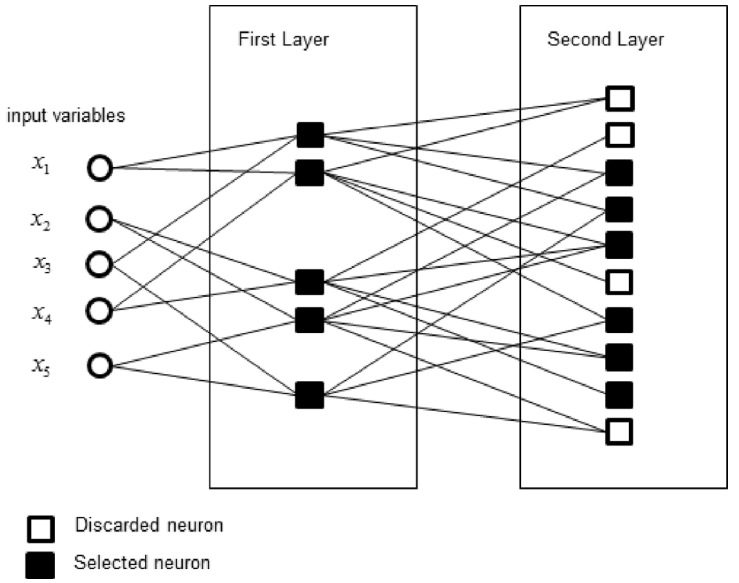

Fig. 3. Second layer of neurons obtained with the multi-layered GMDH modelling process.

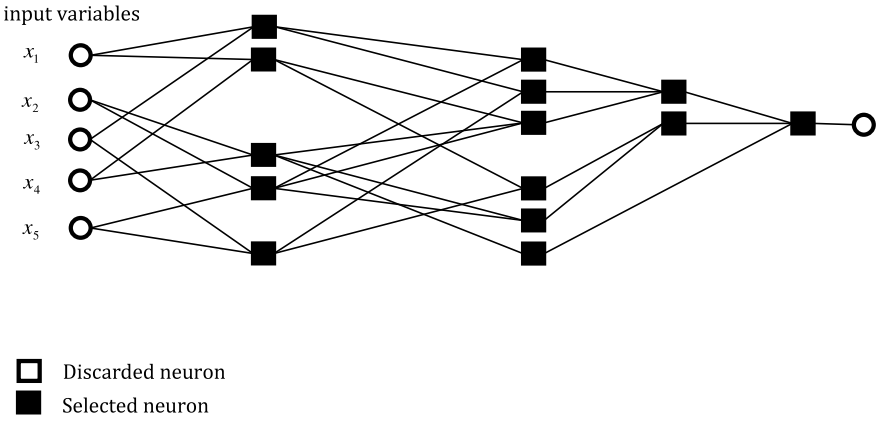

Fig. 4. Final network architecture obtained with the multi-layered GMDH modelling process.

be combined pairwise again:

$$\forall k \in \{2, 3, \ldots\};\ \forall j \in \left\{1, 2 \ldots, \binom{m_{k-1}}{2}\right\};\ \forall i, r \in \{1, 2, \ldots, m_{k-1}\},\ i \neq r, \quad y_j^k = \psi\left(y_i^{k-1}, y_r^{k-1}\right). \quad (15)$$

Thus, through the use of evolution (pairwise combination of parameters and neurons) and selection, more complex organizations are generated subsequently from layer to layer until a final optimally complex model has been obtained that trades-off model accuracy on training data and predictive power on out-of-sample data according to the variance of the dataset (Fig. 4).

Various data splitting rules exist and are used in practice. We have adopted the leave-one-out (LOO) cross-validation for this purpose [1,22], which does not require explicit data subdivision and thus does not unnecessarily shorten already small training datasets. Using an LOO cross-validation method, the original sample is implicitly divided into two parts N times, the training subsample $N_A = N - 1$ and the testing subsample $N_B = 1$. The first subsample is used to estimate the coefficients of a certain gradually more complex polynomial function (hypothesis) while the testing subsample is used to validate the hypothesis thus steering the evolution of the structure of the opti-

mally complex model subsequently from layer to layer. In each layer the best models are selected by the minimum of the external criterion value.

Note that, unlike classical artificial neural networks (ANN), there is no need to predefine the number of neurons or layers to be used as these are adaptively determined through the learning process. The key difference between classical ANN and GMDH is the self-organization property which proceeds through an *inductive* 'bottom-up' approach (model complexity deterministically increases in an adaptive fashion) rather than a *deductive* 'top-down' approach in which the network structure is postulated a priori (sometimes followed by a simplification step referred as 'pruning' according to a priori selected parameters). Unlike most ANN approaches where the structure is predefined (fixed network structure with parameters estimated through the optimization of a highly multimodal surface), the optimal structure estimation including its corresponding explicit mathematical expression is a basic part of the GMDH iterative procedure, which is key for the proposed HDMR expansion approach.

Also, there is no stopping rule to be set a priori, since the model self-organization stops itself when model optimal complexity has been found. Optimality in this regard means that further increasing model complexity would result in over-fitting the design data by starting to adapt to the structural uncertainty (i.e. the model would tend to use unessential parameters or bilinear interaction for the fitting without recognizing the structural uncertainty that exists because interactions of more than two parameters are not considered). For these reasons, despite the use of networks of neurons the approach presented in this paper differs from other ANN emulators based approaches to calculated HDMR expansions such as in [7].

3.2. An algorithm for using GMDH to compute an HDMR expansion

In this section we present the step-wise method which relies on the direct construction of the RS-HDMR expansion through GMDH inductive modelling:

Step 1: Considering a set of parameters $(x_i)_{i\in[\![1,M]\!]}$ taking values in the hypercube $[0,1]^M$ (scaling original parameters in $[0,1]^M$ if necessary), 'synthetic' parameters $X_{p,i}$ are built using the Legendre orthogonal polynomials (Fig. 5): $X_{p,i} = \varphi_p(x_i)$, $p\epsilon\mathbb{N}^*$ where $p \in \mathbb{N}^*$ represent the polynomial order for the shifted Legendre family of basis functions which have the following form:

$$\begin{aligned}
\varphi_1(x_i) &= \sqrt{3}(2x_i - 1),\\
\varphi_2(x_i) &= 6\sqrt{5}\left(x_i^2 - x + \frac{1}{6}\right),\\
\varphi_3(x_i) &= 20\sqrt{7}\left(x_i^3 - \frac{3}{2}x_i^2 + \frac{3}{5}x - \frac{1}{20}\right)\\
&\vdots\\
\varphi_p(x_i) &= (-1)^p\sqrt{2p+1}\sum_{k=0}^{p}\binom{p}{k}\binom{p+k}{k}(-x_i)^k\\
&\vdots
\end{aligned} \tag{16}$$

These 'synthetic' parameters are evaluated only once in this step for all the realizations and are used subsequently by the GMDH expansion building process.

Step 2: Define elementary functions of neurons using the following bilinear representation:

$$\begin{aligned}
&\forall p,\ q \in \mathbb{N}^{*2},\ p < q,\ \forall i, j = \{1,\ldots,M\},\ i < j,\\
&\psi_{i,j}^{p,q}\left(\varphi_p(x_i), \varphi_q\left(x_j\right)\right) = \psi_0 + a\varphi_p(x_i) + b\varphi_q\left(x_j\right) + c\varphi_p(x_i)\varphi_q\left(x_j\right)\\
&\qquad\qquad = \psi_0 + aX_{p,i} + bX_{q,j} + c\cdot X_{p,i}\cdot X_{q,j},
\end{aligned} \tag{17}$$

where ψ_0 is a constant term and a, b, c are constant scalar coefficients. The neurons $\psi_{i,j}^{p,q}$ correspond to pairs of polynomials φ_p and φ_q evaluated on original parameters x_i and x_j, excluding products between polynomials evaluated on the same original parameter (i.e. $i \neq j$).

Step 3: Finding an expression for the model now consists of finding the optimal subset of parameters $\alpha \subseteq \{1, M\}$, $|\alpha| \leq 2$ and associated decomposition functions f_α, i.e. $\{f_\alpha\}_{\alpha\subseteq\{1,M\},\ |\alpha|\leq 2}$, the set of polynomials

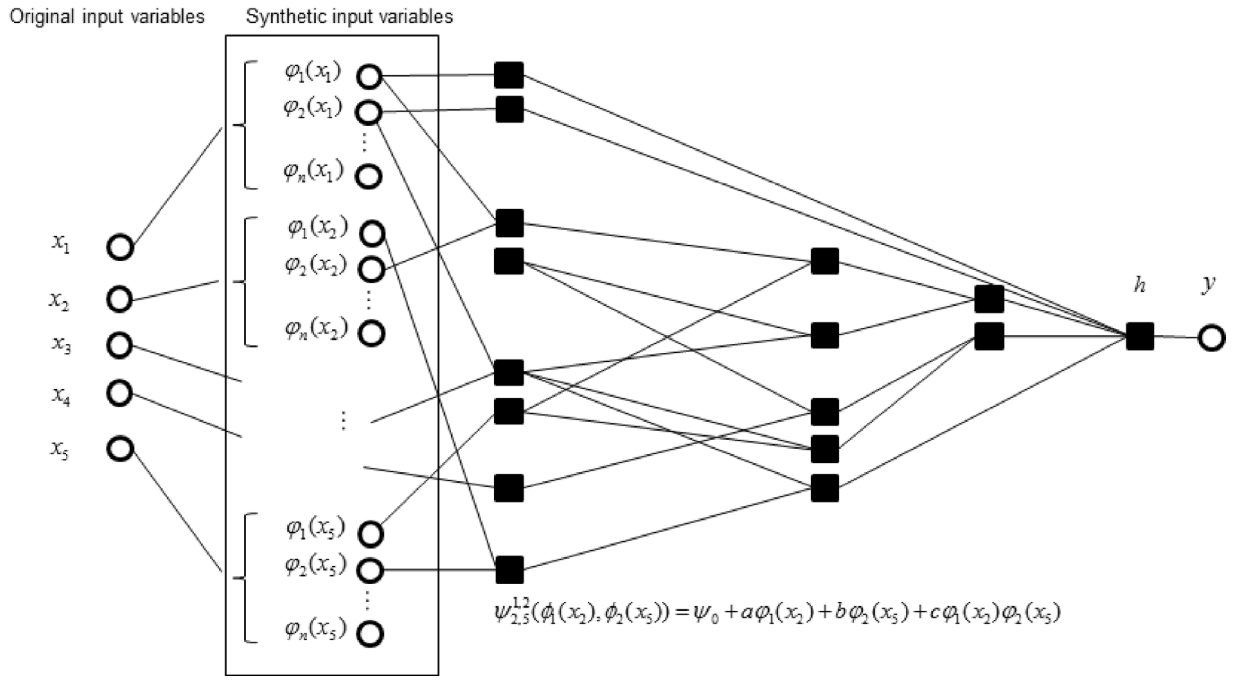

Fig. 5. GMDH network showing original and synthetic parameters used in deriving the HDMR expansion.

$\{\varphi_p\}_{p\in u\subseteq\mathbb{N}^*}$ and the values of coefficients $\{\alpha_r^i\}_{r\in\mathbb{N}^*,i\in\alpha}$ and $\left\{\beta_{pq}^{ij}\right\}_{p,q\in\mathbb{N}^*,i,j\in\alpha}$ in (9). This step corresponds to an optimization whose decision parameters are not only coefficients in the decomposition using predefined polynomial orders but also the polynomial orders and the subset of important parameters.

The subsequent layers of the GMDH network are formed by allowing any combination of neurons which results in a multilinear or linear expression of the synthetic parameters. This step is performed through the use of the multi-layered iterative GMDH algorithm presented in the previous section.

Remark. A 'synthetic parameter' corresponds to evaluations of shifted Legendre polynomial functions on the original parameters which have been scaled to $[0, 1]^M$.

Step 4: The coefficients in the final GMDH expression are then used to calculate the Sobol sensitivity indices according to formulas (5), (12) and (13).

The GMDH modelling approach produces a structure and parameter estimation problem which leads to the discarding of:

(a) All synthetic parameters $X_{p,i}$ corresponding to the Legendre polynomial basis function of the original unessential parameter x_i;
(b) All synthetic parameters $X_{p,i}$ corresponding to suboptimal polynomial order p for a given retained original parameter x_i. This corresponds to the selection of the optimal polynomial order for the HDMR representation. The problem of optimal polynomial order has been discussed in a number of studies [2,31,32];
(c) Unessential/suboptimal products of Legendre polynomials $\varphi_p(x_i)\ \varphi_q\left(x_j\right)$ that do not significantly contribute to the overall variance of the system.

4. Application examples

4.1. The Ishigami function

In this example, we consider the so-called Ishigami function, a classical benchmark function for GSA methods [23,25], which is a highly nonlinear function of three inputs:

$$f(x) = \sin(\pi(2x_1 - 1)) + 7(\sin(\pi(2x_2 - 1))^2) + 0.1\pi^4(2x_3^4 - 1)\sin(\pi(2x_1 - 1)), \tag{18}$$

Table 1
Comparison of Sobol sensitivity indices obtained from GMDH with analytical results for the Ishigami function.[a]

	Analytical	GMDH-HDMR ($N = 64$)	GMDH-HDMR ($N = 128$)	GMDH-HDMR ($N = 256$)	GMDH-HDMR ($N = 512$)
S_1	0.3139	0.261	0.291	0.3186	0.313
S_2	0.4424	0.442	0.439	0.4428	0.442
$S_{1,3}$	0.2437	0.275	0.248	0.236	0.242

[a] Values of sensitivity indices in applications examples of Section 4 have been calculated using a C++ implementation of the presented algorithm. Commercial versions of the software are available at http://www.knowledgeminer.eu/ockham/.

where $x_i,\ i = 1, 2, 3$ are uniformly distributed on the interval [0, 1]. Comparison of the results of the GMDH-HDMR approximation with the analytically computedvalues of Sobol sensitivity indices are presented in Table 1.

One noticeable effect of the GMDH inductive modelling is the increase in the number of α_i and $\beta_{i,j}$ coefficients (increase in number of selected polynomial orders) along with an increase of the sample size. Unlike classical regression, GMDH does not compute the coefficients corresponding to every polynomial up to a certain order, but selects the important contributions in an evolutionary fashion. In this example the GMDH-HDMR approximation with 512 sample points contains three more terms in comparison to the expansion obtained with $N = 256$. The expressions in (20) and (21) give an account of selected polynomial orders for each parameter ($\varphi_p(x_i)$ is the shifted Legendre polynomial of order p for parameter x_i). Formulas are given in variables scaled to the unit interval [0,1]:

GMDH-HDMR expansion for $N = 256$:

$$\begin{aligned} Y = {} & 1.63392\varphi_1(x_1) - 1.3167\varphi_3(x_1) - 0.613232\varphi_2(x_2) - 2.01623\varphi_4(x_2) + 1.29603\varphi_6(x_2) \\ & + 1.3617\varphi_1(x_1)\varphi_2(x_3) - 0.993099\varphi_3(x_1)\varphi_2(x_3) + 3.50991. \end{aligned} \tag{19}$$

GMDH-HDMR expansion for $N = 512$:

$$\begin{aligned} Y = {} & 1.622414\varphi_1(x_1) - 1.30349\varphi_3(x_1) + 0.18688\varphi_5(x_1) - 0.59350\varphi_2(x_2) - 1.95261\varphi_4(x_2) \\ & + 1.359423\varphi_6(x_2) - 0.34924\varphi_8(x_2) + 1.387746\varphi_1(x_1)\varphi_2(x_3) - 1.09496\varphi_3(x_1)\varphi_2(x_3) \\ & + 0.4073377\varphi_1(x_1)\varphi_4(x_3) + 3.50996. \end{aligned} \tag{20}$$

The GMDH-HDMR approximation avoids the computational error that stems from deriving the value of all potential coefficients of the HDMR expansion. It highlights the ability of GMDH to automatically select the optimal polynomial order according to the size of the sample. In effect, as explained before and contrary to other techniques, GMDH construction is not only a parameter estimation task but also a structure optimization procedure. Hence it allows for the selection of the optimal polynomial order as a function of sample size. It is consistent with observations made in the case of the use of RS-HDMR, where higher polynomial orders typically entail a higher number of sample points to calculate the coefficients with sufficient accuracy [25]. In the next section we test the methodology on well-known high dimensional test functions.

4.2. The Sobol g-200 function

In this example we test the approach on the so-called Sobol g-function, which is often used in assessing sensitivity analysis techniques. It has the following form [23]:

$$f(x) = \prod_{i=1}^{M} \frac{|4x_i - 2| + a_i}{1 + a_i}. \tag{21}$$

For this function the variance D of $f(x)$ and the Sobol sensitivity indices can be computed analytically as follows: Here the input parameters $x_i,\ i = 1, \ldots, M$ are uniformly distributed over $[0, 1]^M$ and the

$$D = \prod_{i=1}^{M} (D_i + 1) - 1 \tag{22}$$

$$D_i = \frac{1}{3(1 + a_i)^2} \tag{23}$$

Table 2
Comparison of the values of Sobol sensitivity indices obtained with GMDH-HDMR with analytical values for the g-function with 200 parameters ($M = 200$).

	Analytical	GMDH-HDMR, $N = 256$	GMDH-HDMR, $N = 512$	GMDH-HDMR, $N = 1024$	GMDH-HDMR, $N = 2048$
S_1	0.225	0.232	0.226	0.221	0.224
S_2	0.141	0.144	0.143	0.145	0.142
S_3	0.100	0.127	0.104	0.0963	0.102
S_4	0.073	N/A	0.077	0.0726	0.0721
S_5	0.0562	N/A	0.0562	0.0597	0.0557
S_6	0.025	N/A	0.0273	0.0318	0.0255
S_7	0.014	N/A	0.0108	0.014	0.015
S_8	0.009	N/A	N/A	0.00716	0.0099
S_9	0.0062	N/A	N/A	N/A	0.00592
$S_{1,2}$	0.048	0.0382	0.0322	0.0425	0.0385
$S_{1,3}$	0.033	0.0367	0.0453	0.026	0.0268
$S_{1,4}$	0.0245	N/A	0.0178	0.0198	0.0196
$S_{1,5}$	0.018	N/A	0.0246	0.0248	0.0125
$S_{1,6}$	0.008	N/A	N/A	0.0169	0.00574
$S_{1,7}$	0.00469	N/A	N/A	N/A	0.00435
$S_{2,3}$	0.0213	0.0455	N/A	0.0227	0.0195
$S_{2,4}$	0.01568	N/A	N/A	N/A	0.0114
$S_{2,5}$	0.0120	N/A	N/A	N/A	0.0078
$\sum_i \sum_j (S_i + S_{ij})$	0.935	0.624	0.765	0.795	0.807

$$S_{i_1,\ldots,i_s} = \frac{1}{D} \prod_{i=1}^{s} D_i \tag{24}$$

coefficients a_i are non negative. The lower the value of coefficient a_i, the more significant parameter x_i.

In this example we choose $M = 200$ with the coefficients a_i such that $\{a_1, \ldots, a_{20}\} = [0, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]$, $\forall i > 20$, $a_i = 99$.

This particular instance of the Sobol g-function with 200 parameters was presented in the work of Touzani and Busby [27]. The GMDH-HDMR method was applied using different numbers of samples as shown in Table 2. One can see that the number of selected parameters increases with an increase of the sample size allowing for the computing of more sensitivity indices. $N = 2048$ samples are sufficient for rather accurate approximations of significant sensitivity indices. Adding more samples only marginally improves the results.

4.3. The K-function

This function was previously used in [14,21] as an illustrative example for variance-based sensitivity analysis techniques. It contains only a few dominant variables and mainly low order interactions (i.e. it has a small effective dimension in the truncation sense [14]). The function has the following form:

$$f(x) = \sum_{i=1}^{M} (-1)^i \prod_{j=1}^{i} x_j. \tag{25}$$

Vector x is uniformly distributed on $[0, 1]^M$. The analytical values for the main effect sensitivity indices can be found in [21].

In this paper we use an instance of this function with $M = 100$. From Table 3, it can be seen that GMDH-HDMR is able to select the correct important parameters and that the values of total Sobol sensitivity indices are close to the predicted analytical values, although with a more significant discrepancy for $S^f_{T_4}$ for which only S_4 was computed. This can be due to implementation and tolerance issues and the fact that interactions of three or more parameters may be needed to account for the total variance of the function.

Table 3
Comparison of the values of the total Sobol sensitivity indices obtained with GMDH-HDMR with analytical values for the K-function with 100 parameters ($M = 100$).

Total index	Analytical	GMDH-HDMR ($N = 256$)	Structure of $S^f_{T_i}$
$S^f_{T_1}$	0.75	0.74	$S_1 + S_{1,2} + S_{1,3}$
$S^f_{T_2}$	0.25	0.23	$S_2 + S_{1,2}$
$S^f_{T_3}$	0.08	0.06	$S_3 + S_{1,3}$
$S^f_{T_4}$	0.03	0.01	S_4

Table 4
Comparison of the values of Sobol sensitivity indices obtained with GMDH-HDMR with analytical values for the Rosenbrock function with 10 parameters ($M = 10$) for $N = 256$.

$S^f_{T_i}$	Analytical values	GMDH-HDMR	$S^f_{T_i}$ structure for GMDH-HDMR
$S^f_{T_1}$	0.085	0.084	$S_1 + S_{1,2}$
$S^f_{T_2}$	0.177	0.163	$S_2 + S_{1,2} + S_{2,3}$
$S^f_{T_3}$	0.177	0.161	$S_3 + S_{2,3} + S_{3,4}$
$S^f_{T_4}$	0.177	0.171	$S_4 + S_{3,4} + S_{4,5}$
$S^f_{T_5}$	0.177	0.168	$S_5 + S_{4,5} + S_{5,6}$
$S^f_{T_6}$	0.177	0.169	$S_6 + S_{5,6} + S_{6,7}$
$S^f_{T_7}$	0.177	0.176	$S_7 + S_{6,7} + S_{7,8}$
$S^f_{T_8}$	0.177	0.173	$S_8 + S_{7,8} + S_{8,9}$
$S^f_{T_8}$	0.177	0.179	$S_9 + S_{8,9} + S_{9,10}$
$S^f_{T_9}$	0.100	0.101	$S_{10} + S_{9,10}$

4.4. The Rosenbrock function

This example is another well-known benchmark function for which total sensitivity indices can be calculated analytically.

$$f(x) = \sum_{i=1}^{M-1} (x_i - 1)^2 + 100\left(x_i^2 - x_{i+1}\right)^2. \tag{26}$$

For $M = 10$ and 256 samples we obtained the following results presented in Table 4:

These results once again demonstrate the efficiency of the proposed approach for functions with low order interactions.

5. Conclusions

A new method for the calculation of the Sobol global sensitivity indices has been presented. This method is based on the use of GMDH for the derivation of the component functions of HMDR. The method has three main advantages: (1) the possibility to deal with high dimensions especially in the case of under-determined tasks; (2) the ability to perform with a limited number of function evaluations (useful in the case of expensive model computations); (3) a capability for optimal selection of parameters, polynomial orders and parameter selections of the high dimensional model representations.

The developed methodology has been tested on well-known model functions to illustrate the benefits of simultaneous inductive structure identification and parameter optimization of meta-models. This approach is similar to the adaptive techniques developed by Blatman and Sudret [2,3] since both techniques are adaptive and use cross-validation. The key difference is that the approach by Blatman and Sudret constructs the metamodel as a *linear* regression of polynomials

$$h(x) = \sum_{\alpha \subseteq \{1,\ldots,M\}} f_\alpha(x) = \sum_{\alpha \subseteq \{1,\ldots,M\}} a_\alpha \varphi_\alpha(x) = a^T \varphi(x), \tag{27}$$

where for $\alpha = \{p, q\}$, $\varphi_\alpha = \varphi_{pq}\left(x_i, x_j\right) = \varphi_p(x_i)\,\varphi_q\left(x_j\right)$ is considered as a separate additional parameter built from individual components $\varphi_p(x_i)$ and $\varphi_q\left(x_j\right)$ in the regression. In the approach used in this paper, the multi-linear formulation of the problem allows for the consideration of the product of polynomials without the need to consider their product as an additional separate parameter (thereby avoiding an increase in the number of parameters for the regression). We note that HDMR is a particular case of Polynomial Chaos Expansion (PCE) [2,3], which can be seen as truncated PCE to low order interactions in the ANOVA decomposition. In principle, there is nothing that prevents the use of GMDH to compute a full PCE. However, for practical implementation reasons we have limited ourselves to low order interactions. Higher interactions will be considered in future research.

Acknowledgement

The financial support of the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 314441 (CELSIUS) is gratefully acknowledged.

References

[1] D. Allen, The prediction sum of squares as a criterion for selecting prediction variables, Technical Report 23, Department of Statistics, University of Kentucky, 1971.
[2] G. Blatman, B. Sudret, Efficient computation of global sensitivity indices using sparse polynomial chaos expansions, Reliab. Eng. Syst. Saf. 95 (11) (2010) 1216–1229.
[3] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Phys. 230 (6) (2011) 2345–2367.
[4] E. Borgonovo, E. Plischke, Sensitivity analysis: A review of recent advances, European J. Oper. Res. 248 (3) (2016) 869–887.
[5] R. Chowdhury, S. Adhikari, High dimensional model representation for stochastic finite element analysis, Appl. Math. Model. 34 (12) (2010) 3917–3932.
[6] S. Dey, T. Mukhopadhyay, S. Adhikari, Stochastic free vibration analysis of angle-ply composite plates—a RS-HDMR approach, Compos. Struct. 122 (2015) 526–536.
[7] W. Hao, Z. Lu, P. Wei, J. Feng, B. Wang, A new method on ANN for variance based importance measure analysis of correlated input variables, Struct. Saf. 38 (2012) 56–63.
[8] T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of nonlinear models, Reliab. Eng. Syst. Saf. 52 (1) (1996) 1–17.
[9] B. Ioos, P. Lemaitre, in: G. Dellino, C. Moloni (Eds.), Uncertainty Management in Simulation-Optimization of Complex Systems, Springer, 2015, pp. 101–122.
[10] A.G. Ivakhnenko, Polynomial theory of complex systems, IEEE Trans. Syst. Man Cybern. SMC-1 (4) (1971) 364–378.
[11] A.G. Ivakhnenko, J.-A. Müller, Self-organization of nets of active neurons, Syst. Anal. Modelling Simul. 20 (1–2) (1995) 93–106.
[12] A.G. Ivakhnenko, J.-A. Müller, GMDH algorithms for complex systems modelling, Math. Comput. Model. Dyn. Syst. 4 (1998) 275–316.
[13] T.L. Kelley, An unbiased correlation ratio measure, Proc. Natl. Acad. Sci. USA 21 (9) (1935) 554–559.
[14] S. Kucherenko, B. Feil, N. Shah, W. Mauntz, The identification of model effective dimensions using global sensitivity analysis, Reliab. Eng. Syst. Saf. 96 (4) (2011) 440–449.
[15] G. Li, S.-W. Wang, H. Rabitz, Practical approaches to construct RS-HDMR component functions, J. Phys. Chem. A 106 (37) (2002) 8721–8733.
[16] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of Sobol indices for the Gaussian process metamodel, Reliab. Eng. Syst. Saf. 94 (3) (2009) 742–751.
[17] J.-A. Müller, F. Lemke, Self-organizing modelling and decision support in economics, Syst. Anal. Modelling Simul. 18–19 (1995) 135–138.
[18] J.E. Oakley, A. O'Hagan, Probabilistic sensitivity analysis of complex models: a Bayesian approach, J. R. Stat. Soc. Ser. B Stat. Methodol. 66 (3) (2004) 751–769.
[19] H. Rabitz, Ö.F. Aliş, J. Shorter, K. Shim, Efficient input–output model representations, Comput. Phys. Comm. 117 (1–2) (1999) 11–20.
[20] A. Saltelli, P. Annoni, How to avoid a perfunctory sensitivity analysis, Environ. Model. Softw. 25 (12) (2010) 1508–1517.
[21] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, S. Tarantola, Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, Comput. Phys. Comm. 181 (2) (2010) 259–270.
[22] G.A.F. Seber, A.J. Lee, Linear Regression Analysis, second ed., Wiley, 2003.

[23] I.M. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Math. Comput. Simul. 55 (1–3) (2001) 271–280.
[24] V.S. Stepashko, Method of critical variances as analytical tool of theory of inductive modeling, J. Autom. Inf. Sci. 40 (3) (2008) 4–22.
[25] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliab. Eng. Syst. Saf. 93 (7) (2008) 964–979.
[26] J. Tissot, C. Prieur, Bias correction for the estimation of sensitivity indices based on random balance designs, Reliab. Eng. Syst. Saf. 107 (2012) 205–213.
[27] S. Touzani, D. Busby, Screening method using the derivative-based global sensitivity indices with application to reservoir simulator, Oil Gas Sci. Technol. 69 (4) (2014) 619–632.
[28] H. Wang, L. Tang, G.Y. Li, Adaptive MLS-HDMR metamodeling techniques for high dimensional problems, Expert Syst. Appl. 38 (11) (2011) 14117–14126.
[29] P. Wei, Z. Lu, J. Song, Variable importance analysis: A comprehensive review, Reliab. Eng. Syst. Saf. 142 (2015) 399–432.
[30] H. Xiong, Z. Chen, H. Qiu, H. Hao, H. Xu, Adaptive SVR-HDMR metamodeling technique for high dimensional problems, AASRI Procedia 3 (2012) 95–100.
[31] T. Ziehn, A.S. Tomlin, Global sensitivity analysis of a 3D street canyon model—Part I: The development of high dimensional model representations, Atmos. Environ. 42 (8) (2008) 1857–1873.
[32] T. Ziehn, A.S. Tomlin, GUI–HDMR—a software tool for global sensitivity analysis of complex models, Environ. Model. Softw. 24 (7) (2009) 775–785.