DOI: 10.3901/CJME.2016.0422.058, available online at www.springerlink.com; www.cjmenet.com

Data Driven Uncertainty Evaluation for Complex Engineered System Design

LIU Boyuan[1,*], HUANG Shuangxi[1], FAN Wenhui[1], XIAO Tianyuan[1], James HUMANN[2], LAI Yuyang[3], and JIN Yan[2]

1 *State CIMS Engineering Research Center, Tsinghua University, Beijing 100084, China*
2 *University of Southern California, Los Angeles, California, USA*
3 *SOYOTEC Technologies Co., Ltd., Beijing 100081, China*

Received November 27, 2015; revised March 4, 2016; accepted April 22, 2016

Abstract: Complex engineered systems are often difficult to analyze and design due to the tangled interdependencies among their subsystems and components. Conventional design methods often need exact modeling or accurate structure decomposition, which limits their practical application. The rapid expansion of data makes utilizing data to guide and improve system design indispensable in practical engineering. In this paper, a data driven uncertainty evaluation approach is proposed to support the design of complex engineered systems. The core of the approach is a data-mining based uncertainty evaluation method that predicts the uncertainty level of a specific system design by means of analyzing association relations along different system attributes and synthesizing the information entropy of the covered attribute areas, and a quantitative measure of system uncertainty can be obtained accordingly. Monte Carlo simulation is introduced to get the uncertainty extrema, and the possible data distributions under different situations is discussed in detail. The uncertainty values can be normalized using the simulation results and the values can be used to evaluate different system designs. A prototype system is established, and two case studies have been carried out. The case of an inverted pendulum system validates the effectiveness of the proposed method, and the case of an oil sump design shows the practicability when two or more design plans need to be compared. This research can be used to evaluate the uncertainty of complex engineered systems completely relying on data, and is ideally suited for plan selection and performance analysis in system design.

Keywords: complex engineered system design; uncertainty; data-driven evaluation; Monte Carlo simulation

1 Introduction

Today's highly developed science and technology have made engineered systems more complex than ever before. This complexity extends the development cycle time and increases the development cost. Systems engineers have been developing effective system design techniques for decades, such as the well-known Quality Function Deployment (QFD)[1], Unified Program Planning (UPP)[2], Axiomatic Design Method (ADM)[3] and Design Structure Matrix (DSM)[4], to mention a few. One common feature of these methods is the assumption of the interdependencies between possible design parameters, components, or tasks can be identified so that they can be either excluded (e.g., by applying the independence axiom in Axiomatic Design) or managed (e.g., by using a design structure matrix). For the situations where such interdependencies may not be known to designers and other unintended interactions may emerge, new methods are needed that can help designers analyze the uncertainty of the system being designed and predict the impact of their design decisions on system uncertainty.

For complex systems, the sheer number of highly correlated variables makes the internal relations of a system hard to understand and manage. Any unexpected disturbance may cause a dramatic change to the system. The more complex the system is, the larger the uncertainty of the system is, as complexity significantly affects a system's chances of fulfilling its functional requirements[5]. Being unaware of this inherent uncertainty in a system may result in a fragile system. Because this uncertainty is caused by the inherent characteristics of complex systems, traditional modelling methods may not be suitable. We propose a data driven uncertainty evaluation approach to support complex engineering design by providing a data-mining based uncertainty evaluation method to assist design decision-making. The method is composed of three main phases as shown in Fig. 1: data preparation, data processing and result analysis. The data to be evaluated should be collected in a data warehouse in advance. The data can be simulation data, experimental data, or any other data that need to be evaluated. The main steps of the data preparation phase include data transformation, which transforms the unrecognizable data to standard form; data

* Corresponding author. E-mail: liuboyuan@126.com
Supported by National Hi-tech Research and Development Program of China (863 Program, Grant No. ##)

cleaning, which filters the outliers that will skew the result; and data reduction, which will reduce the size to be evaluated and increase the speed of the evaluation.

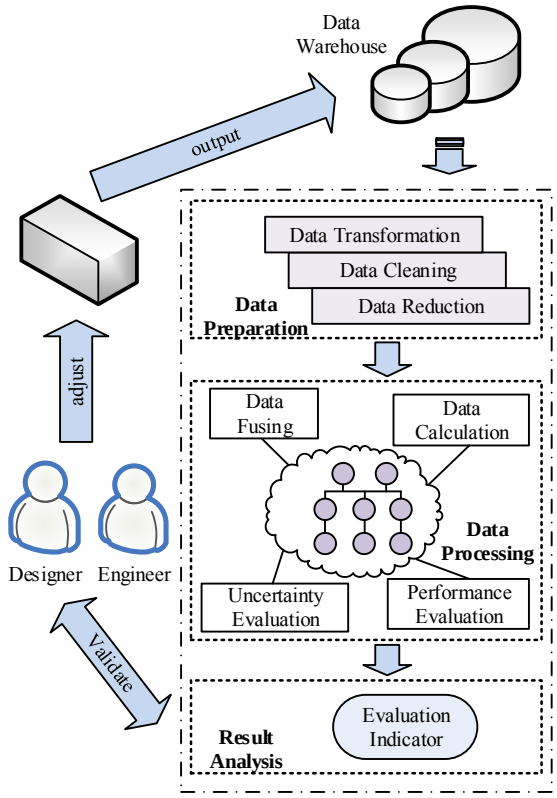

Fig. 1. A data-driven uncertainty evaluation approach

By combining different arithmetical operations and data handling methods, the data processing phase can generate the uncertainty metric and other performance indicators. In the phase of result analysis, the evaluation results are given to designers and engineers. The output at the end of the evaluation will result in system requirements for the subsequent detailed design phase. It is acknowledged that at regular intervals within the design process it is necessary for the state of the design solution to be evaluated[6], which will provide guidance for further design or iterations. Designers can adjust the system according to the evaluation results.

In this paper, we focus on the "uncertainty evaluation" box in Fig. 1 and introduce general data driven uncertainty metrics to fulfil the need of rapid uncertainty evaluation in complex engineered systems. The remainder of this paper is organized as follows. In section 2, the related work is briefly reviewed. Section 3 proposes a novel data processing method to measure system uncertainty, and section 4 describes a Monte Carlo simulation based uncertainty value normalization. Two case studies are discussed in section 5, and conclusions and future work are presented in section 6.

2 Related Work

Much research has been done in the areas of complex systems, data-mining and uncertainty quantification. The fields of study have been motivated by two modern trends: an increase in the complexity of engineered systems, and an increase in data sensing, storage, and processing capabilities. In this section, we briefly survey the studies of complex systems, data mining, and uncertainty quantification within the context of engineering design, with a motivation and theoretical base in complexity theory.

Complex systems have many parts whose interactions cause difficulty in analysis. These interactions can stem from dependencies based on geometry, function, or transfer of energy and material[7]. Complexity can also present in different forms during design, relating to the design problem, the design process, or the designed artifact[8]. Complexity can cause problems such as uncertainty and unclear or nonlinear relationships between a system's inputs and outputs[9]. The consequence of complexity is not just subjective difficultly perceived by designers, but increased design effort and cost, as SINHA, et al[10] link a quantitative measure of structural complexity (based on the connectivity matrix of component interactions) to real-world consequences such as increased assembly time in experimental subjects and tentatively propose a nonlinear dependence of system development budget on system complexity. Another complexity quantification theory and its derivative software, OntoSpace, can quantify complexity by working with unfiltered data and applying a holistic quantitative score to the health of a system or company[11–12]. Designers have used several strategies to overcome the uncertainty caused by complexity. Uncertainty-based robust design optimization is widely used to reduce the effects of uncertainty during design[13–14]. CHALUPNIK, et al[15], compared different "ilities", such as flexibility and reliability that can make a system insensitive to uncertainty.

Engineers already have a long history of exploiting structured and well-organized data. Knowledge bases exist to aid designers in quickly retrieving relevant data as their design work progresses. Researchers introduced a failure mode knowledge base that connects with functional descriptions to aid in failure mode and effects analysis in early design phases[16]. Similar research has built a repository of successful designs focused on their function, behavior, and structure[17]; a biologically-inspired design aid[18]; and even preliminary steps toward the automated synthesis of functional systems based on a data set of functions and linkage constraints[19]. The effort that goes into building these knowledge bases is often quite extensive, and could be considered more as an academic pursuit than a profitable business operation.

As a next step, data mining can be used to gather knowledge from data sets that are much easier to generate, but also less organized. Data is information in its rawest form. It can only be classified as information or knowledge once it has been cognitively filtered[20]. With the increasing

power of computers, the increasing affordability of data storage, and the increasing connectivity and ubiquity of sensors comes the ability to generate vast amounts of data, and there are calls to transform this data into a competitive advantage[21]. Data can be generated in-situ by products or through simulation. Some legacy data may already exist on company servers but is unused[22]. To interpret vast data stores automatically, many firms turn to data mining. KUSIAK, et al[23], outline areas of product and manufacturing system design for data mining applications in design such as pattern recognition and prediction of customer behavior. ROMANOWSKI, et al[24], apply data mining to support variant design activities. KIM, et al[25], present a data-mining-aided optimal design method which is able to find a competitive design solution with a relatively low computational cost. The potential applications are so great that, for modern complex systems, data generation and mining should be included in considerations of a product's total lifecycle[26].

Uncertainties found in complex engineered systems can be results of both parametric and structural uncertainties. They can also be aleatoric or epistemic uncertainties. From a design point of view, structural uncertainty and epistemic uncertainty are potentially risky and can be the sources of system failures. They should therefore be avoided or reduced. Researchers in the area of uncertainty quantification (UQ) have proposed various methods for assessing and managing uncertainties including forward uncertainty propagation[27–30], sensitivity analysis[31–32], response surface tools[33–34], and dimensional reduction tools[35]. These stochastic data analysis methods have been very useful in assessing and modeling uncertainty based on limited availability of data. When modeling capability is limited and the amount of available data is large, a data-mining based approach is more effective. Recently, data-mining techniques have started to play an important role in uncertainty quantification[36].

As technology progresses, the expectations placed on designers will only increase, and they will be faced with designing systems that are more and more complex. This complexity is manifest in uncertainty about system performance, increased design effort, and difficulty in manufacturing. A simultaneous increase in the capabilities of data collection and analysis, however, is providing resources that engineers can use to overcome complexity and uncertainty. Through the application of data mining, engineers can make use of abundant data to gain knowledge about their systems, and increase their confidence in their performance. Our research aims to provide a data-mining based uncertainty evaluation tool for engineers to understand the impact of their design decisions on the uncertainty of the resulting system.

3 Uncertainty Evaluation

Uncertainty is an inherent property of systems. It characterizes the amount of "noise" in the information flow[37]. This kind of noise is usually measured by entropy. Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Classical data mining methods can typically discover such knowledge as association rules, clustering rules, classification rules, and regression rules. As data handling tools are developed, and processing speeds rise, the range of knowledge that data mining can bring to us also extends. Here we propose a novel data mining method which will assign an uncertainty metric to an engineered system. The process of uncertainty evaluation is summarized in Fig. 2, with details given in the following sections.

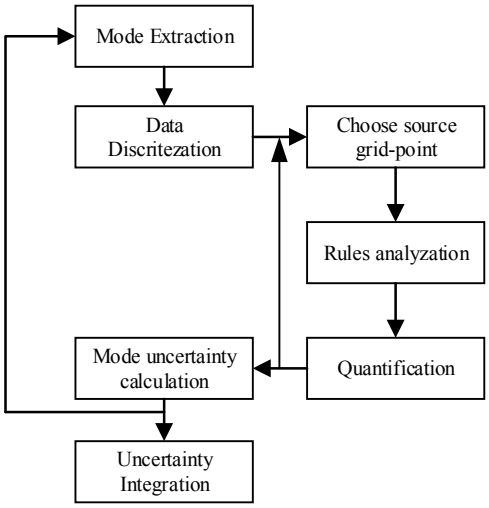

Fig. 2. Process of uncertainty evaluation

What we should prepare for the analysis is only data, which can be simulation data or operational data. The initial data should be organized in the structure of a bivariate table. For given n samples with m attributes, the set of the initial data can be represented as

$$\boldsymbol{D} = [A_1 \quad A_2 \quad \cdots \quad A_m] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix},$$

where A_i is a specific attribute of data; a_{ij} is the concrete value corresponding to the attribute; it must be numeric.

3.1 Extracting modes

The first step is to organize valid modes. We choose different attributes as a benchmark in turn, and analyze each attribute's relationship with another attribute by scatterplot. For 3 attributes, we can get 6 scatterplots: A_1-A_2, A_1-A_3, A_2-A_1, A_2-A_3, A_3-A_1 and A_3-A_2. It can be seen that for m attributes, we can get $m(m-1)$ scatterplots. For scatterplot A_1-A_2, A_1 is taken as X-axis and A_2 is taken as

Y-axis. Across every axis, the lower bound is the minimum value of its corresponding attribute, and the upper bound is the maximum value.

The dataset we use contains continuous variables, and set-mining techniques have been largely designed for categorical or discrete data, so the common method of dealing with these is to discretize them by breaking them into ranges[38]. The concept of evaluation precision is introduced to ameliorate several problems associated with the fact that almost all data values are different. Evaluation precision indicates the granularity at which we deal with the data, and the level of precision determines how many intervals the data will be divided into. In light of the data size, evaluation precision is usually 5 to 9, but no more than 10. Higher evaluation precision may bring more accurate results, but will also lead to higher computational complexity.

The most popular methods of discretization are equal-width, equal-frequency, K-means and maximal entropy. Because the data from our example is continuous data of one specific index, and it will not have sharp discontinuities, here the equal-width method is used. The example of scatterplot A_1-A_2 is shown above the arrow in Fig. 3. The dataset is created artificially so that the principle of the algorithm can be described clearly. In a real application, this could be an example of an engine's temperature plotted vs power output. In Fig. 3 the evaluation precision is set to 5, which forms 5×5 grids.

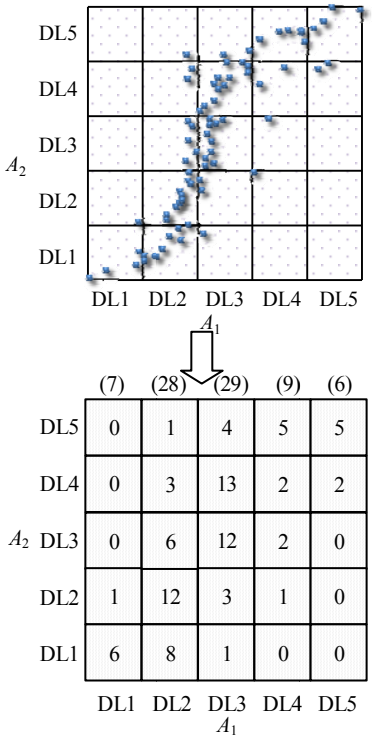

Fig. 3. Scatterplot and mode of A_1-A_2

After discretization, the precise original values are not important anymore because within every grid-point they are considered the same. So the scatterplot of A_1-A_2 can be transformed to another form which is called "mode" as shown below the arrow in Fig. 3. The numbers in mode represent how many points are there in the grid-point. All extracted modes need to be analyzed and we will mine possible association patterns in them.

Because the process of discretization will not be entirely accurate, whether a point belongs to a certain grid-point cannot just be judged by its absolute position. If one point is within a 5% data level length of the boundary, it belongs to both grid-points. For grid-point {1, 1}, the two points in grid-point {2, 1} near the boundary within 5% will be considered to belong to {1, 1}, so the number in {1, 1} is 6. The numbers in parentheses above modes represent the total number of points in the data level they belong to. Sometimes the number does not equal to the sum of its lower numbers because of repeat counts in different grid-points. To facilitate the analysis, the different regions are named from data level (DL) 1 to 5 on both axes.

3.2 Analyzing modes

Here we analyze the correlation between different attributes. We do this by applying a small disturbance on the mode and analyzing the possible change due to the disturbance. The strength of this disturbance should be sufficient to make the value vary for one level (positive or negative). If the disturbance is not big enough to change its value, the attribute is considered unchanged and there is no need to discuss. What we try to mine is some rules in a form similar to the following:

$$X + \Delta X \to \Delta Y(\Delta X = \pm 1DL), \tag{1}$$

Why must we analyze the situation in which an attribute varies only one level? Because every attribute is continuous, and even if it changes quickly, the change is continuous. What does the rule represent? If a disturbance occurs on one attribute, whether from internal or external factors, its value must change. The rule indicates the reaction of another attribute due to this variation. Furthermore, all the rules in a system indicate the change brought on by disturbances, and constitute the uncertainty of a system.

The extraction of this rule is similar to the mining process of an association rule. The most important concepts in association rules are confidence and support. Before mining, a grid-point should be chosen as the source. The definitions of confidence and support used in our analysis are as follows.

Confidence: The proportion of all points on this data level contained in the source grid-point.

Support: After being disturbed, the proportion of all points on that data level contained in one grid-point.

The minimum threshold is set to:

$$\min\_\text{conf} = 2/\text{EP},\ \min\_\text{sup} = 2/\text{EP},$$

where EP is the evaluation precision.

The average proportion of points in one grid-point on one data level is $1/\mathrm{EP}$, and the threshold here is twice as much as the average value. This threshold value, set by the designer, determines how many rules we can get. A higher value can avoid many useless rules. For the evaluation precision of 5, the value is 40%. When the source grid-point reaches min_conf, it can be used to mine possible rules. If there are fewer than 3 points on one data level, the grid-points on this level are all discarded, to avoid mining trivial rules. When a certain grid-point reaches min_sup for a rule, the rule is established. So it is obvious that not every mode will yield a rule.

In Fig. 4, the grid-point {2, 2} is chosen as the source grid-point, and the data count in it is 12. On data level 2, the total count of data is 28. So the confidence is 12/28=42.9%. After a positive disturbance has been applied to attribute A_1, the value of A_1 will jump from data level 2 to data level 3. All the possible grid-points that the attribute may vary to on data level 3 are {3, 1}, {3, 3}, {3, 4} and {3, 5}, and their support levels are 3.4%, 41.4%, 44.8% and 6.9%. Grid-point {3, 2} is not taken into account because even if it reaches min_sup, it means that the change A_1 will not affect A_2 as they are on the same data level. As {3, 3} and {3, 4} both reach minimum support, we have reason to believe that after some positive disturbance, attribute A_2 is likely to vary to data level 3 or data level 4. In Fig. 4, a line with an arrow is depicted to show the trend. So the association rules on grid-point {2, 2} we mined can be formulated as follows:

Rule1: $X+\Delta X \rightarrow 1DL(\Delta X=+1DL)$,
[support=41.4%, confidence=42.9%].

Rule2: $X+\Delta X \rightarrow 2DL(\Delta X=+1DL)$,
[support=44.8%, confidence=42.9%].

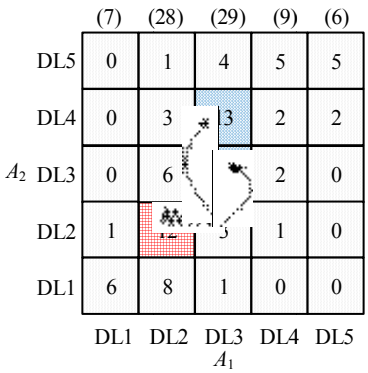

Fig. 4. Positive disturbance of A_1-A_2

If the evaluation precision is high, there will be more rules which satisfy min_sup that can be mined. Only two rules which can cover the maximum area will be accepted.

It is the same with negative disturbances. We apply a negative disturbance to attribute A_1, and it varies from data level 2 to data level 1. On data level 1, the support of grid-point {1, 1} is 85.7%. So the association rule mined is:

Rule 3: $X+\Delta X \rightarrow -1DL(\Delta X=-1DL)$,
[support=85.7%, confidence=42.9%].

The complete rule is shown in Fig. 5. The mode contains at least one rule, so it is called a valid mode.

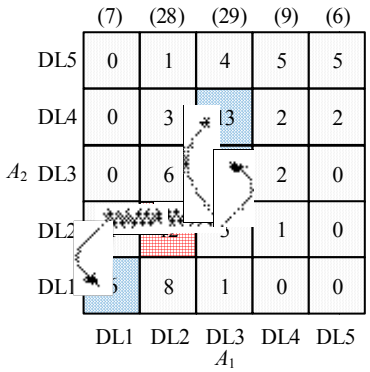

Fig. 5. Complete mode of A_1-A_2

3.3 Computing uncertainty of a mode

Now we have the association rule in grid-point {2, 2} with A_1 as the benchmark. Let us focus on the subarea that involves the association rule. This area is shown in Fig. 6.

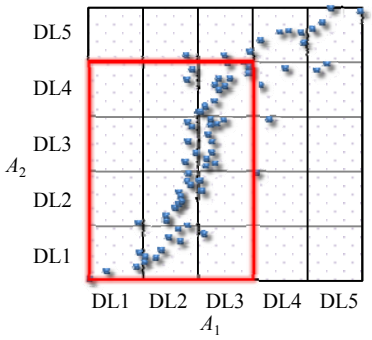

Fig. 6. Subarea of mode

In this area, the correlation between data is more obvious. Next, information entropy is used to characterize the strength of this correlation. Shannon's information entropy is applied here to describe how much information is carried by the association rules or how strong the correlation is. Information entropy is a measure of the uncertainty of a random variable. Eq. (2) gives the expression of information entropy:

$$H=-\sum_{i=1}^{n} p(x_i)\log p(x_i), \tag{2}$$

Just as in data mining, to get the rational information entropy value, these values also need be discretized. Here, an equidistant partition is used, and the region will be divided into $\lfloor\sqrt{n}\rfloor\times\lfloor\sqrt{n}\rfloor$ grids, where n is the data count in the region. For the subarea shown in Fig. 6, the points in this area is 56, which means the region will be

divided into 7×7 grids as shown in Fig. 7.

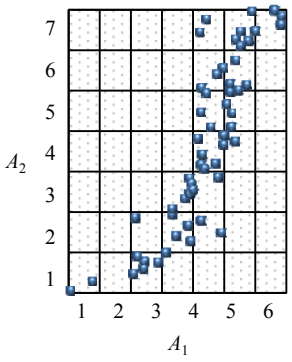

Fig. 7. Subarea of mode

The distribution of different grid-points is shown in Table 1, which can be used to calculate the information entropy. In Table 1, "points" means how many points are contained in one grid-point, "probability" means the probability that the grid-point will appear in Fig. 7, and "numbers" means how many grid-points of this type are in Fig. 7.

Table 1. Probability distribution of the subarea

Parameter	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
Points	1	2	3	4	6	7
Probability	0.0179	0.0357	0.0536	0.0714	0.1071	0.125
Numbers	6	3	3	4	2	1

So the information entropy of this area is

$$\begin{aligned} H_{2,2} = -(&6 \times 0.0179 \times \log 0.0179 + \\ &3 \times 0.0357 \times \log 0.0357 + 3 \times 0.0536 \times \log 0.0536 + \\ &4 \times 0.0714 \times \log 0.0714 + 2 \times 0.1071 \times \log 0.1071 + \\ &1 \times 0.125 \times \log 0.125) = 2.751. \end{aligned}$$

Here the base of the logarithm is e, and the unit of entropy is nat. Because the entropy is generated by choosing grid-point {2, 2} as the source, we mark it with $H_{2,2}$. The smaller the entropy is, the stronger the relevance is, meaning that the uncertainty of this mode is smaller.

For one scatterplot, every grid-point can be chosen as the source grid-point. So what we should notice is that different subarea containing different rules may have the same information entropy, but they will not perform the same role in a system. The confidence of a grid-point corresponds to the probability that a system may go into that state, so the confidence value we have determined is used to balance its role. Hence the weighted uncertainty of a mode equals:

$$U(A_x, A_y) = \sum_{i=1}^{\mathrm{EP}} \sum_{j=1}^{\mathrm{EP}} \mathrm{confidence}(i, j) \bullet H_{i,j}. \tag{3}$$

3.4 Computing uncertainty of a system

As previously mentioned, for one variable, at most (m–1) modes can be mined. So the uncertainty caused by one attribute is the sum of its uncertainty with different modes. Every variable can be chosen as a disturbance source to analyze possible relevance, so the uncertainty of the whole system is

$$\begin{aligned} U = &\sum_{x=1}^{m} \sum_{y=1, y \neq x}^{m} U(A_x, A_y) = \\ &\sum_{x=1}^{m} \sum_{y=1, y \neq x}^{m} \left(\sum_{i=1}^{\mathrm{EP}} \sum_{j=1}^{\mathrm{EP}} \mathrm{confidence}\,(i, j) \bullet H_{ij} \left| M_{A_x, A_y} \right. \right), \end{aligned} \tag{4}$$

where M_{A_x, A_y} indicates that the calculation is in the mode consisting of A_x and A_y.

4 Uncertainty Normalization

The metrics proposed in section 3 give us a quantitative description of the system uncertainty, but it is obvious that different systems contain different association patterns, the same system in different phases may follow different patterns, and even the same pattern probably involves different data sizes, which will result in different uncertainty values with no direct comparability. How to compare the uncertainty values under different situations becomes the next problem that must be solved.

For different modes, the uncertainty is measured by information entropy, so there should be a maximum and minimum information entropy that a mode can reach. Since the uncertainty of a system is the combination of different modes, there will also be a maximum and minimum uncertainty for the system. After getting the maximum, minimum and current uncertainty of a system, a normalization representation can be used to give a standard value to compare uncertainties in different situations. Now we try to analyze the maximum and minimum values.

As shown in Fig. 6, the area in a mode that involves the association rules can be determined. For two variables, a linear relation is the simplest pattern, and our assumption is that this could be used to determine the minimum value of uncertainty; the relation full of noise is the complicated model, which could be used to determine the maximum value. The following analysis is based on this thought. For this area, the horizontal length is 3 levels (denoted by L_x), and the vertical length is 4 levels (denoted by L_y).

Here, Monte Carlo simulation is introduced to obtain the extrema. For the situation with minimum information entropy, all points are distributed over the linear area with a perfect linear correlation. In fact, it is hard to estimate the true distribution, so we introduce a distribution with added noise. The normal distribution is an extremely important concept in statistics, and is often used in the natural and social sciences for real-valued random variables whose

distributions are not known[39]. So a normal distribution is used here to simulate the mode with minimum entropy, and the same number of points is used.

For the distribution of X, μ is set to $L_x/2$, which means the average value is the center point of the area. σ is set to $L_x/4$. So the probability density function is:

$$f(x)=\frac{1}{\sqrt{2\pi}\frac{L_x}{4}}\exp\left(-\frac{\left(x-\frac{L_x}{2}\right)^2}{2\left(\frac{L_x}{4}\right)^2}\right)= \frac{4}{\sqrt{2\pi}L_x}\exp\left(-\frac{8\left(x-\frac{L_x}{2}\right)^2}{L_x^{\,2}}\right),\ 0<x<L_x. \tag{5}$$

After getting the X simulation value, the ordinate Y can be found by the linear relation:

$$y=\frac{L_y}{L_x}\bullet x. \tag{6}$$

The simulation result is shown in Fig.8.

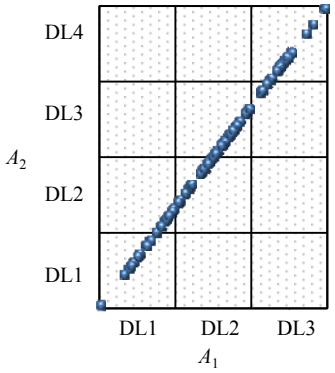

Fig. 8. Scatterplot with minimum possible entropy

This mode can be used to get the minimum possible entropy. As the same number of points is used, the partition of this area for the calculation of information entropy is also the same as the original status, and it will not affect the result. The result of the calculation is

$$H_{2,2}^{\min}=-\sum_{i=1}^{n}p(x_i,y_i)\log p(x_i,y_i)=1.935.$$

The mode with maximum entropy is more complex than the minimum situation. The mode with the maximum possible entropy should have the following characteristics: first, the mode should contain at least one supported association relationship; it cannot only be noise. Second, the mode should be as chaotic as possible, which will affect the identification of the mode.

To satisfy the first demand, the distribution of X is still based on a linear relation, but unlike the previous situation, the linear relation is based on the rules we mined before. For the mode we have been analyzing, the linear relationship is shown in Fig.9 with a blue line.

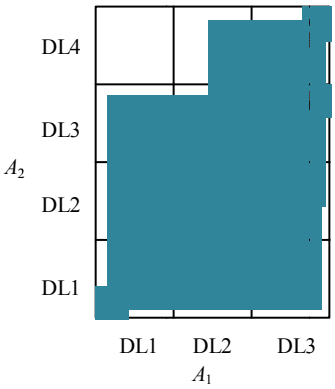

Fig. 9. Diagram of linear relation

A uniform distribution is applied to it (the entropy of a uniform distribution is bigger than the normal distribution). It should be noted that the distribution of X as uniform, does not mean that every interval between all points is equal. The distribution is dependent on simulation results. The probability function of X is

$$f(x)=\frac{1}{L_x},0<x<L_x. \tag{7}$$

To satisfy the second demand, a normal distribution is applied to generate the deviation from the linear approximation. Here μ is set to 0, which means that the average value is on the linear area; σ is set to $12\sqrt{L_x^{\,2}+L_y^{\,2}}/(L_y/L_x)$, which will ensure that the deviation degree is not too excessive. Another uniform distribution is not suitable here because it will make the mode totally chaotic and no more rules will be found. As this normal distribution is perpendicular to the linear area, the transformation of distance to coordinate is more complicated:

$$x=x'+d\sin\left(\arctan\frac{L_y}{L_x}\right), \tag{8}$$

$$y=y'-d\cos\left(\arctan\frac{L_y}{L_x}\right). \tag{9}$$

For one simple linear relation, the scatter simulation with 500 points to test our hypothesis is shown in Fig.10. It can be seen in Fig. 10 that the points are mainly distributed in the linear area, and some noise is shown around the linear

area, which will not affect the identification of a valid mode.

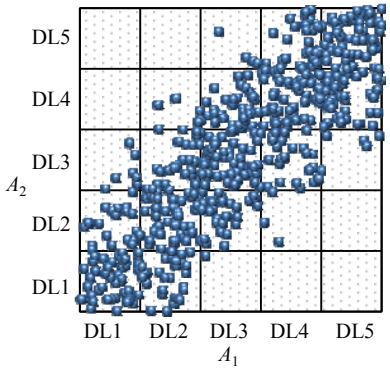

Fig. 10. Linear relation with maximum entropy

For the mode we have been focusing on, the simulation result is shown in Fig. 11.

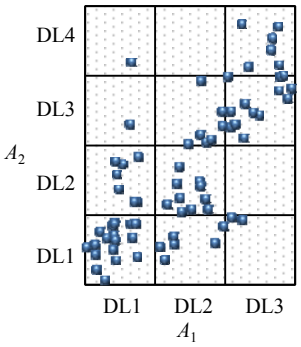

Fig. 11. Scatterplot with maximum possible entropy

As the number of points is small, the mode is not as clear as the relation shown in Fig. 10. The maximum possible information entropy via our simulation is

$$H_{2,2}^{\max} = -\sum_{i=1}^{n} p(x_i, y_i)\log p(x_i, y_i) = 3.108.$$

For the question we are concerned with, we use $H^{\max}$, $H^{\min}$ and H, and can get a normalized value:

$$\tilde{H} = \frac{H - H^{\min}}{H^{\max} - H^{\min}}. \tag{10}$$

For this mode, the normalized value is

$$\tilde{H}_{2,2} = 0.664.$$

For the evaluation of a system, the normalized uncertainty is based on the total maximum and minimum uncertainty:

$$U^{\max} = \sum_{x=1}^{m}\sum_{y=1,y\neq x}^{m}\left(\sum_{i=1}^{\mathrm{EP}}\sum_{j=1}^{\mathrm{EP}}\mathrm{confidence}\,(i,j)\,H_{ij}^{\max}\left|M_{A_x,A_y}\right.\right), \tag{11}$$

$$U^{\min} = \sum_{x=1}^{m}\sum_{y=1,y\neq x}^{m}\left(\sum_{i=1}^{\mathrm{EP}}\sum_{j=1}^{\mathrm{EP}}\mathrm{confidence}\,(i,j)\,H_{ij}^{\min}\left|M_{A_x,A_y}\right.\right). \tag{12}$$

So the normalized uncertainty of a system is

$$\tilde{U} = \frac{U - U^{\min}}{U^{\max} - U^{\min}}. \tag{13}$$

5 Case Studies

In this section, two case studies are discussed. The case of an inverted pendulum system is introduced to verify the effectiveness of the proposed approach, and a case of oil sump design is used to show its value in practice. The proposed approach is coded in C++ and runs on a Core i7 3.40 GHz PC.

5.1 An inverted pendulum

An inverted pendulum system consists of a pendulum, a lightweight bar, a base and a spring. The base will execute simple harmonic motion. The system is shown in Fig. 12. The inverted pendulum is a classic example of an inherently unstable system[40], and its nature, "sensitive dependence on initial conditions," causes it to exhibit great uncertainty when running, and makes it suitable to verify our theory.

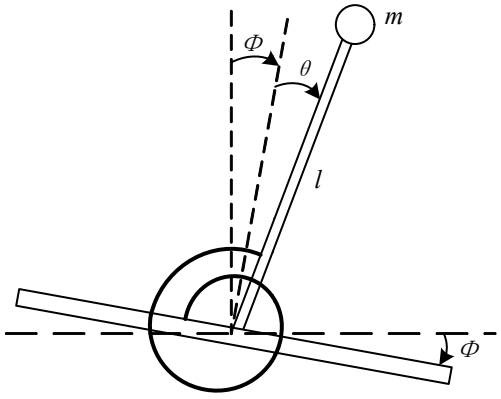

Fig. 12. An inverted pendulum system

The Duffing Equation derived from this system is

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \delta\frac{\mathrm{d}x}{\mathrm{d}t} - x + x^3 = f\cos\omega t, \tag{14}$$

where δ is a dimensionless damping coefficient, ω is dimensionless angular frequency and f is the dimensionless amplitude of the driving force. Different initial conditions with slight dissimilarity are specified to observe their results. 3 different initial conditions are

$$\left[x, \frac{\mathrm{d}x}{\mathrm{d}t}, f, \omega, \delta\right] = \begin{bmatrix} 0.1 & 0.1 & 1 & 1 & 0.78 \\ 0.1 & 0.1+0.001 & 1 & 1 & 0.78 \\ 0.1 & 0.1-0.001 & 1 & 1 & 0.78 \end{bmatrix}.$$

Next the system is simulated in Matlab. The data recorded by Matlab are shown in Table 2, and the plot is shown in Fig. 13. The time and pendulum angle in both Table 2 and Fig 13 are dimensionless.

Table 2. Simulation data of inverted pendulum system

Time t	Angle 1 θ_1	Angle 2 θ_2	Angle 3 θ_3
0	0.1	0.1	0.1
0.01	0.101 051	0.101 041	0.101 061
0.02	0.102 203	0.102 183	0.102 223
0.03	0.103 456	0.103 427	0.103 486
0.04	0.104 809	0.104 77	0.104 849
0.05	0.106 262	0.106 213	0.106 311
0.06	0.107 813	0.107 754	0.107 872
…	…	…	…
99.96	0.285 913	−1.211 11	0.299 434
99.97	0.289 848	−1.207 19	0.303 422
99.98	0.293 864	−1.203 16	0.307 491
99.99	0.297 96	−1.199 03	0.311 641
100	0.302 137	−1.194 79	0.315 872

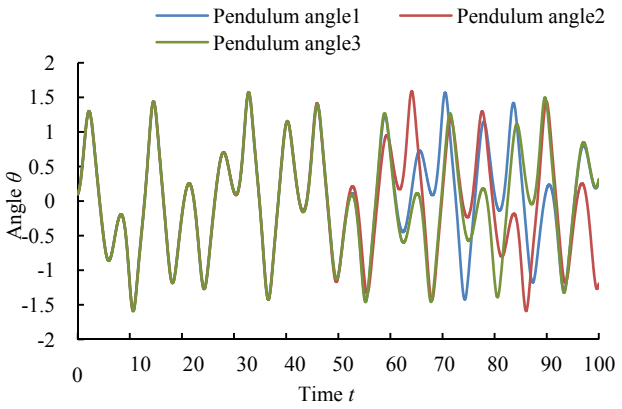

Fig. 13. Pendulum angle curve of inverted pendulum system

It can be seen from Fig. 13 and Table 2, at the initial time, the pendulum angles of all 3 motions are nearly identical. After 50, differences begin to appear. As time goes on, the differences become increasingly obvious. This extreme sensitivity to initial conditions for deterministic equations is usually considered chaos[41]. We treat the system as two separate systems according to the time period, and divide it into two parts: the first 50 and the last 50.

Thus we are in essence measuring the output of two separate systems. Because only one attribute can be measured in the system, 3 angle data under different initial conditions are regarded as three different attributes, and their correlation is analyzed by the proposed method. For this system of about 5000 data points each, the evaluation precision is set to 7. All the evaluation results are as follows.

Uncertainty of first 50:

$$U^{\mathrm{cur}} = 248.882,\ U^{\mathrm{min}} = 216.326,$$
$$U^{\mathrm{max}} = 393.024,\ \tilde{U} = 18.42\%.$$

Uncertainty of last 50:

$$U^{\mathrm{cur}} = 183.634,\ U^{\mathrm{min}} = 117.409,$$
$$U^{\mathrm{max}} = 207.591,\ \tilde{U} = 73.43\%.$$

From the evaluation results, it is obvious that the uncertainty value of the last 50 is bigger than the uncertainty of the first 50, which indicates the uncertainty brought on by a transition into chaos. By linking rising uncertainty, measured by our quantitative metric, to a known example of a system transitioning into chaos, this case study verifies the validity of the uncertainty metrics successfully.

5.2 An oil sump design

In order to evaluate its actual role in practice, an oil sump design is taken as an empirical case. As an important part of diesel engines, an oil sump is used to store lubricant and seal the engine, and also helps in heat dissipation.

The model of the oil sump is created in HyperMesh. Two plans are put forward and modeled separately. Most of the dimensions are the same for both plans, and the parameters that vary between plans are shown in Table 3. These parameters are explained by the labels in Fig. 14 and Fig. 15.

Table3. Parameters in two designs

Parameter	Plan 1 (mm)	Plan 2 (mm)
Back_Reinforcing	6.0	5.0
Bottom_R	5.4	4.0
Front_Reinforcing1	5.0	5.0
Front_Reinforcing2	6.0	5.0
Left_Reinforcing1	2.0	5.0
Left_Reinforcing2	2.0	5.0
Left_Reinforcing3	2.0	5.0
Left_Reinforcing4	2.0	5.0
Right_Reinforcing1	2.0	5.0
Right_Reinforcing2	2.0	5.0
Right_Reinforcing3	2.0	5.0
Side_Wall_R	4.1	4.0
Upper_R	12.0	8.0

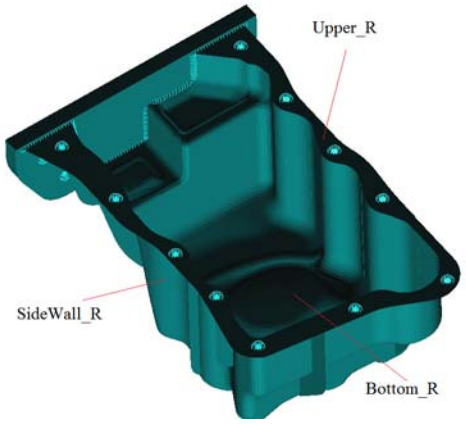

Fig. 14. Top-view parameters of oil sump

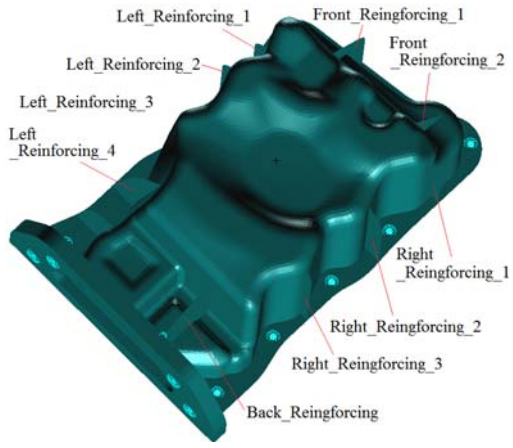

Fig. 15. Bottom-view parameters of oil sump

For the oil sump design, there are many important performance metrics such as rigidity, strength and modal shape. Here its dynamic characteristics are considered by analyzing resonance in a frequency domain. At a resonance point, intense vibration responses will occur. The vibration response can be reduced by changing the structure to avoid a resonance point, or by improving rigidity. Next the two plans are analyzed in Nastran. Nastran will output the acceleration of different viewpoints on the oil sump as a response to an oscillating disturbance. The movement of the viewpoints will represent the overall vibration performance, and the designer can analyze the performance according to the data. The attributes are as follows:

- *Y* acceleration of viewpoint #1 on bottom;
- *Y* acceleration of viewpoint #2 on bottom;
- *Y* acceleration of viewpoint #3 on bottom;
- *Y* acceleration of viewpoint #4 on bottom;
- *Y* acceleration of viewpoint #5 on bottom;
- *X* acceleration of viewpoint #1 on back wall;
- *X* acceleration of viewpoint #2 on back wall;
- *X* acceleration of viewpoint on front wall;
- *Z* acceleration of viewpoint on right wall;
- *Z* acceleration of viewpoint on left wall.

To get enough data to analyze the two plans completely, every parameter in Table 3 fluctuates independently within 10% to get the simulation results. Some of the data generated by plan 1 are shown in Table 4.

Table 4. Simulation data of plan 1

Run	*Y* acc. on bottom #1 $A1/(\mathrm{m}\bullet\mathrm{s}^{-2})$	…	*Z* acc. on right wall $A9/(\mathrm{m}\bullet\mathrm{s}^{-2})$	*Z* acc. on left wall $A10/(\mathrm{m}\bullet\mathrm{s}^{-2})$
1	6.296 9	…	149.311 7	49.644 51
2	6.563 6	…	3127.138	864.195 6
3	6.2	…	257.988	62.283 02
4	6.478 7	…	1086.718	609.315 4
5	6.090 9	…	58.485 92	114.114 1
…	…	…	…	…
100	6.442 4	…	146.263 7	251.592 2

The two plans both can meet the high-level design requirements well, and therefore as an auxiliary index, our uncertainty metric is introduced to evaluate them.

For this system of 100 data samples each, the evaluation precision is set to 5. For plan 1, analyzed by the proposed algorithm, 46 modes are extracted. In all the extracted modes, there are 9 modes that take the *X* acceleration of point #1 on the back wall as a source, and 11 modes that take the *X* acceleration of point #2 on the back wall as a source. Because the two attributes create the most associations with others, they are the most important attributes in the plan. After synthesizing all modes, we can get the following results.

Uncertainty of plan 1:

$$U^{\mathrm{cur}} = 37.7918,\ U^{\mathrm{min}} = 29.798\,8,$$
$$U^{\mathrm{max}} = 49.439\,8,\ \tilde{U} = 40.72\%.$$

For plan 2, 63 modes are extracted. The most important attributes in plan 2 are the *Y* acceleration of point #3 on the bottom, which holds 10 modes, and the *X* acceleration of the viewpoint on the front wall, which holds 8 modes. The results are as follows.

Uncertainty of plan 2:

$$U^{\mathrm{cur}} = 37.416\,5,\ U^{\mathrm{min}} = 26.263\,9,$$
$$U^{\mathrm{max}} = 44.059\,2,\ \tilde{U} = 62.67\%.$$

From the evaluation results, it is obvious that the uncertainty value of plan 2 is larger than that of plan 1, which means the oil sump designed according to plan 1 will show higher stability. This also shows that the behavior patterns of plan 2 are more chaotic than plan 1. While both plans can meet the design requirements, for long operation, the oil sump designed by plan 2 may cause more problems such as fatigue or performance degradation.

In practical engineering, for two or more plans to be compared, the plan with the lowest uncertainty value tends to be more reliable. Uncertainty is an essential factor that a designer must account for when deciding among competing design options, given that all variants meet the original project requirements. During the design process of a system, if the modification of one factor makes the uncertainty index higher, it means the reliability of the system will be lowered, and the designers should be alert to this tradeoff and adjust the system accordingly.

6 Conclusions

(1) The uncertainty metrics proposed in this paper provide a new perspective to analyze and understand a system, and give us a quantitative index to describe the nature of its uncertainty, which is especially useful for developing complex engineered systems.

(2) The analysis process of the proposed method is completely based on data, which decreases the requirement of detailed internal knowledge of a system. By means of

mining association relations along different system attributes, and quantifying the relations using information entropy, a quantitative uncertainty index can be obtained.

(3) Monte Carlo method is introduced to simulate the possible uncertainty extrema under different situations, and a normalized uncertainty quantitative index is presented accordingly which enhances the usability of the proposed method.

(4) The two case studies have demonstrated the effectiveness and practicality of the method. When faced with complex systems, design engineers can use this method to assess the uncertainty impact of their design decisions and increase their odds of fulfilling the system's functional requirements.

(5) Several steps can be taken to improve upon the basis established here. In this paper, the data distribution for the uncertainty extrema was based on hypothesis, but we plan to develop a more objective distribution to provide a precise estimation. We also plan to apply the metrics and approach to more practical case studies beyond the oil sump, and attempt to quantitatively determine a correlation between a system's uncertainty and real-world design metrics such as design time, design cost, and product reliability.

References

[1] AKAO Y. Development history of quality function deployment[J]. *The Customer Driven Approach to Quality Planning and Deployment*, 1994: 339.

[2] HILL J D, WARFIELDJN. Unified program planning[J]. *Systems, Man and Cybernetics, IEEE Transactions on*,1972(5): 610–621.

[3] SUHN P. *The principles of design*[M]. Oxford University Press, 1990.

[4] STEWARD D V. *Systems analysis and management: structure, strategy and design*[M]. New York: Petrocelli Books, 1981.

[5] SUH N P. A theory of complexity, periodicity and the design axioms[J]. *Research in Engineering Design*, 1999, 11(2): 116–132.

[6] GREEN G. Modelling concept design evaluation[J]. *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing*, 1997, 11(3): 211–217.

[7] SOSA M E, EPPINGERS D, ROWLES C M. Identifying modular and integrative systems and their impact on design team interactions[J]. *Journal of Mechanical Design*, 2003 125(2): 240–252.

[8] SUMMERS J D, SHAH J J. Developing measures of complexity for engineering design[C]//*ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Chicago, USA, 2003: 381–392.

[9] MINA A A, BRAHA D, BAR-YAM Y. Complex engineered systems: a new paradigm[M]//*Complex Engineered Systems*. Springer Berlin Heidelberg, 2006: 1–21.

[10] SINHA K, DE WECK O L. Structural complexity quantification for engineered complex systems and implications on system architecture and design[C]//*ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Portland, USA, 2013: V03AT03A044-V03AT03A044.

[11] MARCZYKJ, DESHPANDEB. Measuring and tracking complexity in science[M]//*Unifying Themes in Complex Systems*. Springer Berlin Heidelberg, 2008: 27–33.

[12] LOMARIO D, DE POLI GP,FATTORE L. A complexity-based approach to robust design and structural assessment of aero engine components[C]//*ASME Turbo Expo 2007: Power for Land, Sea, and Air*. Montreal, Canada, 2007: 1091–1099.

[13] RAZA M A, LIANG W. Uncertainty-based computational robust design optimisation of dual-thrust propulsion system[J]. *Journal of Engineering Design*, 2012, 23(8): 618–634.

[14] MOHIDEEN M J, PERKINS J D, PISTIKOPOULOS E N. 1996. Optimal design of dynamic systems under uncertainty[J]. *AIChE Journal*, 1996, 42(8): 2251–2272.

[15] CHALUPNIK M J, WYNN D C, CLARKSON P J. Comparison of ilities for protection against uncertainty in system design[J]. *Journal of Engineering Design*, 2013, 24(12): 814–829.

[16] STONE R B, TUMER I Y, VAN W M. 2005. The function-failure design method[J]. *Journal of Mechanical Design*, 2005, 127(3): 397–407.

[17] BOHMM R, STONER B, SIMPSON T W, et al. Introduction of a data schema to support a design repository[J]. *Computer-Aided Design*, 2008, 40(7): 801–811.

[18] NAGEL J K S, STONE R B, MCADAMS D A. An engineering-to-biology thesaurus for engineering design[C]//*ASME 2010 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Montreal, Canada, 2010: 117–128.

[19] BRYANT C R, STONE R B, MCADAMS D A, et al. Concept generation from the functional basis of design[C]//*ICED 05 International Conference on Engineering Design*, Melbourne, Australia, 2005: 1702–1715.

[20] VAN HORN D, OLEWNIK A, LEWIS K. Design analytics: capturing, understanding, and meeting customer needs using big data[C]//*ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Chicago, USA, 2012: 863–875.

[21] SETHII K. Data mining: an introduction[M]//*Data Mining for Design and Manufacturing*, Springer US, 2001: 1–40.

[22] SRIVASTAVA A, HACKER K, LEWIS K, et al. A method for using legacy data for meta model-based design of large-scale systems[J]. *Structural and Multidisciplinary Optimization*, 2004, 28(2): 146–155.

[23] KUSIAK A, SMITH M. Data mining in design of products and production systems[J]. *Annual Reviews in Control*, 2007, 31(1): 147–156.

[24] ROMANOWSKI C J, NAGI R. A data mining approach to forming generic bills of materials in support of variant design activities[J]. *Journal of Computing and Information Science in Engineering*, 2004, 4(4): 316–328.

[25] KIM P, Ding Y. Optimal engineering system design guided by data-mining methods[J]. *Technometrics*, 2004 47(3): 336–348.

[26] ROMANOWSKI C J, NAGI R. A data mining-based engineering design support system: a research agenda[M]//*Data Mining for Design and Manufacturing*, Springer US, 2001: 161–78.

[27] SRINIVASAN R. *Importance sampling: applications in communications and detection*[M]. Springer Science & Business Media, 2013.

[28] BUCKLEW J. *Introduction to rare event simulation*[M]. Springer Series in Statistics, 2013.

[29] SEPAHVAND K, MARBURG S, HARDTKE HJ. Uncertainty quantification in stochastic systems using polynomial chaos expansion[J]. *International Journal of Applied Mechanics*, 2010, 2(2): 305–353.

[30] MOROKOFF W J, CAFLISCH R E. Quasi-Monte Carlo integration[J]. *Journal of Computational Physics*, 1995, 122(2): 218–230.

[31] HELTON J C, DAVIS F J. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems[J]. *Reliability Engineering & System Safety*, 2003, 81(1): 23–69.

[32] SALTELLI A, RATTO M, ANDRES T, et al. *Global sensitivity*

analysis: the primer[M]. John Wiley & Sons, 2008.
[33] KHURI A I, MUKHOPADHYAY S. Response surface methodology[J]. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010, 2(2): 128–149.
[34] HELTON J C, DAVIS F J. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems[J]. *Reliability Engineering & System Safety*, 2003, 81(1): 23–69.
[35] KULLBACK S, LEIBLER R A. On information and sufficiency[J]. *The annals of mathematical statistics*, 1951: 79–86.
[36] KAMATH C. On the Role of Data-mining techniques in uncertainty quantification[J]. *International Journal of Uncertainty Quantification*, 2012, 2(1): 73–94.
[37] SIPPLE H, MARCZYK J. *Application strategies of robust design &complexity management in engineering: current status &future trends in multi-disciplinary product development*[M]. München, WOK Kreuzer, 2009.
[38] BAY S D. Multivariate discretization for set mining[J]. *Knowledge and Information Systems*, 2001, 3(4): 491–512.
[39] CASELLA G, BERGER R L. 2001. *Statistical inference*[M]. Pacific Grove, CA, Duxbury, 2002.
[40] ANDERSON C W. Learning to control an inverted pendulum using neural networks[J]. *Control Systems Magazine*, 1989, 9(3): 31–37.
[41] WANG G Y, YUAN F. Cascade chaos and its dynamic characteristics[J]. *Acta Physica Sinica*, 2013, 62(2): 020506.

Biographical notes

LIU Boyuan, born in 1985, is currently a PhD candidate at *State CIMS Engineering Research Center, Tsinghua University, China*. He received his bachelor degree from *Harbin Institute of Technology, China*, in 2008. His research interests include big data and data mining.
E-mail: liuboyuan@126.com

HUANG Shuangxi, is associate professor at *Department of Automation, Tsinghua University, China*. He received his PhD degree from *Nanjing University of Science and Technology, China*, in 1999. During 2010–2011, he worked as a visiting scholar in aerospace and mechanical engineering at *University of Southern California, USA*. His research interests include collaborative product design, complex system modeling and analysis, and service technology & engineering. He has published more than 40 papers in academic journals and international conferences.

FAN Wenhui, is a professor at *Department of Automation, Tsinghua University, China*. He received his B.S. degree from *Northwestern Poly-technical University, China*, in 1990, and the M.S. degree from *Harbin Institute of Technology, China*, in 1995, and the PhD degree from *Zhejiang University, China*, in 1998. His research interests include collaborative simulation and multidisciplinary optimization.

XIAO Tianyuan, is a professor at *State CIMS Engineering Research Center, Tsinghua University, China*. He received his bachelor's degree from *Tsinghua University, China*, in 1970. His research interests include virtual manufacturing, computer integrated manufacturing systems (CIMS), and cyber-physical system (CPS).

James Humann, is a PhD candidate at *University of Southern California, USA*. His research focuses on machine learning and the design of complex and self-organizing systems

LAI Yuyang, the founder and CEO of SOYOTEC LIMITED, has over ten years of experience of complexity management, computer aided engineering and design optimization. He currently works with global companies including aerospace, automotive and off-shore industries helping them measure and reduce their business complexity and optimize the performance and reliability of their products. He and his team have invented a software for model credibility validation and calibration based on complexity measurement and optimization algorithms.