IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 18, 2020, accepted January 2, 2021, date of publication January 5, 2021, date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049412

A Robust Method to Measure the Global Feature Importance of Complex Prediction Models

XIAOHANG ZHANG[1], **(Member, IEEE), LING WU**[1], **ZHENGREN LI**[2], **AND HUAYUAN LIU**[3]
[1]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]School of Modern Posts, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3]China North Vehicle Research Institute, Beijing 100072, China

Corresponding author: Ling Wu (wuling@bupt.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFC0830801.

ABSTRACT Because machine learning has been widely used in various domains, interpreting internal mechanisms and predictive results of models is crucial for further applications of complex machine learning models. However, the interpretability of complex machine learning models on biased data remains a difficult problem. When the important explanatory features of concerned data are highly influenced by contaminated distributions, particularly in risk-sensitive fields, such as self-driving vehicles and healthcare, it is crucial to provide a robust interpretation of complex models for users. The interpretation of complex models is often associated with analyzing model features by measuring feature importance. Therefore, this article proposes a novel method derived from high-dimensional model representation (HDMR) to measure feature importance. The proposed method can provide robust estimation when the input features follow contaminated distributions. Moreover, the method is model-agnostic, which can enhance its ability to compare different interpretations due to its generalizability. Experimental evaluations on artificial models and machine learning models show that the proposed method is more robust than the traditional method based on HDMR.

INDEX TERMS Feature importance, global interpretation, high-dimensional model representation, robustness, supervised machine learning.

I. INTRODUCTION

Machine learning has been widely used in various fields. For example, in predicting credit scores and health status, machine learning algorithms are used to construct models to map many features into a class (outcome or decision) by a learning process on the digital traces of people's daily activities [1]. Practical requirements often evaluate machine learning models by their accuracy. The pursuit of predictive accuracy leads to the use of more complex predictive models. Simple and interpretable models often do not have the best performance in terms of predictive accuracy [2]. Complex machine learning models, however, are difficult for humans to understand their internal working mechanisms and decision-making process and are commonly referred to as ''black boxes'', such as deep neural networks. Such a lack of transparency can increase severe issues and hinder further applications of machine learning. Conversely, the reason why certain simple models, such as logistic regression and decision tree models, are widely used is partly attributable to the fact that they can generate interpretable models [3]. The interpretability of the black box will increase the trust of users and allow them to realize that it is always possible to understand the decisions made by models [4]. Therefore, for further applications of complex machine learning models, it is necessary to interpret ''black box'' models and make them transparent to builders and users.

Although interpreting models' internal mechanisms and predictive results is a hot topic in the field of machine learning [5], the interpretability of complex models on biased data remains a difficult problem. One inherent risk in the interpretability of complex models is that model users may inadvertently make incorrect decisions with explanations of biased data generated by human or systematic factors. Particularly in risk-sensitive fields, such as self-driving vehicles and healthcare, it is crucial to provide a robust interpretation of complex models when the important explanatory features on concerned data are highly influenced by contaminated distributions. In these scenarios, one incorrect decision may directly lead to death [1]. Considering the interpretability of machine learning models is often associated with analyzing the input features of models by looking

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Masini.

at feature importance. Therefore, providing robust feature importance for input features becomes indispensable in the interpretability of complex models.

Feature importance is an efficient quantitative measure to explore the structure of a model, evaluate the model's response to changes in the model inputs [6], and describe how important the feature is for the predictive performance of the model regardless of the shape (linear or nonlinear relationship) or direction of the feature effect [7]. In the central area of uncertainty in risk assessment, feature importance methods can be used to identify the most critical and essential contributors to output uncertainties and risk [8]. To date, many methods of measuring feature importance have been proposed. We can summarize these feature importance methods from two perspectives: 1) global vs. local and 2) agnostic vs. specific. First, global methods [9]–[15] evaluate feature importance by considering the entire input feature space; however, local methods [16]–[20] only evaluate a single instance's feature importance. Second, agnostic methods can be used with any machine learning model [17], [21], [22], and specific methods are only adaptable to the interpretation and visualization of specific models [19], [20], [23]–[28]. Although specific methods typically have high computational efficiency, users must interpret different black box models with different specific methods, which may increase the difficulty of operation for nonexpert users. Moreover, specific methods cannot be used directly to compare the interpretation of different models. Because agnostic methods can be used with a variety of models [3], they can enhance the ability of the system to compare different interpretations [21], [29].

Only a few feature importance methods that are simultaneously agnostic and global have been proposed [1]. Therefore, this article aims to propose a novel agnostic method to measure global feature importance based on high-dimensional model representation (HDMR). HDMR is a function decomposition technique that is often used to manage either the performance of experiments or the modeling of chemical/physical systems where there are large numbers of input features [30], [31]. However, HDMR is rarely used to evaluate global feature importance in machine learning models. Moreover, although the feature importance method based on analysis of variance (ANOVA) HDMR is theoretically rigorous [15], it is difficult to estimate its variance-based indices correctly with any degree of robustness [32] when there are outliers in the input variables. Therefore, we propose a robust estimation method when the input variables have tiny errors that follow the contaminated distributions.

This article mainly focuses on how to provide robust interpretations of complex models by implementing an agnostic and global feature importance method. The feature importance indicates to what extent an input feature can influence the output. The main contributions of this research are summarized as follows:

1) The proposed method derived from HDMR yields an improvement in robustness compared to the traditional method based on HDMR when the input features follow contaminated distributions.
2) The proposed method is a novel global feature importance method that is model-agnostic. Moreover, due to its general applicability, the method can compare different interpretations among different artificial or machine learning models.

The structure of this article is as follows. In Section II, we briefly summarize the classification of global feature importance research and review the literature related to variance-based measures. In Section III, we discuss the variance-based method derived by HDMR, propose the novel robust method, and discuss relevant calculation issues. Artificial datasets and complex machine learning models are used to simulate and test the robustness of the proposed method in Section IV, and Section V discusses the conclusions of the study.

II. LITERATURE

In this section, we summarize certain global feature importance methods that can be placed into one of two categories: specific and agnostic methods.

Specific methods are only used in the interpretation of specific machine learning models, such as decision trees and random forests. When using specific methods, users must interpret different black box models with different methods. For additive models, the nomogram, which is a visual method for measuring inputs, was previously used to explain naive Bayes models [23] and linear SVM [33]. For random forest models, feature importance is measured by the decrease in prediction accuracy when input features are permuted [11], [12]. A similar permutation measure method has been used for neural networks, where noise is added to input features [34], [35]. Moreover, Welling *et al.* measured feature importance using a method to decompose trees by splitting features for random forest models [13].

Agnostic methods can be used to interpret any model. Certain methods have been proposed for feature sensitivity analysis in engineering. Nonparametric methods using linear regression techniques are the first class of agnostic methods to measure global feature sensitivity [36]–[38]. Then, researchers [14], [15], [39], [40] first used ANOVA-HDMR to decompose interpreted models and derive a variance-based method to measure global feature sensitivity based on the variance of conditional model outputs, assuming that input features are independent. When input variables are dependent, researchers [9], [10], [41], [42] proposed novel methods based on HDMR to investigate global feature sensitivity. Another class of agnostic methods is density-based methods. Researchers use different distance measures to measure the discrepancy between the unconditional model output density and the density conditional on inputs [43], [44]. Moreover, regionalized sensitivity analysis, which is a Monte Carlo filtering procedure, aims to identify which factors are most important in leading to realizations of output that are either in behavioral or non-behavioral regions [45]. Recently, a novel

Shapley feature importance method was proposed to distribute the overall predicted performance of a model fairly among features based on the marginal permutation-based contributions, which can be used to interpret any type of machine learning model [7].

The variance-based method derived by HDMR has been used as a model-agnostic method to measure global feature sensitivity. However, variance is not a robust measure when input features are highly influenced by a contaminated distribution that is associated with the presence of outliers [46]. When estimating variance-based indices, there is a large probability of producing an uncertain feature importance ranking for inputs that are unstable from sample to sample [32], which we address in this study.

III. METHOD

In this section, we first discuss the variance-based method derived by HDMR, then propose the novel robust method, and finally discuss relevant calculation issues.

A. HDMR-BASED FEATURE IMPORTANCE

Given a model of the form $Y = g(X)$, with Y being a scalar output and $X = (X_1, X_2, \ldots, X_n)$ input vector, the variance-based sensitivity analysis is closely related to the decomposition of $g(X)$ into terms of increasing dimensions [47]:

$$Y = g(X) = g_0 + \sum_i g_i(X_i) + \sum_{i<j} g_{ij}(X_i, X_j) + \ldots + g_{12..n}(X_1, X_2, \ldots, X_n) \quad (1)$$

in which each individual term is also square integrable over the domain of existence and is a function only of the factors in its index. This expansion, called high-dimensional model representation (HDMR), is not unique: for a given model $g(x)$, there could be infinite choices for its terms.

Unlike Sobol to derive the global sensitivity indices in ANOVA-HDMR [15], we must not provide the mutually orthogonal condition of all terms in the decomposition (i.e., $\int_X g_{i_1,\ldots,i_s}(X_{i_1}, \ldots, X_{i_s})dX_k = 0$ for any $k = i_1, \ldots, i_s$). We can first let these terms be a particular form, which is unequivocally calculated using conditional expectations of the model output:

$$\begin{aligned} g_0 &= \int_X g(X) f_X(X) dX = \mathbb{E}[g(X)] \\ g_i &= g_i(X_i) = \int_{X_{\sim i}} g(X) f_{X_{\sim i}}(X_{\sim i}|X_i) dX_{\sim i} - g_0 \\ &= \mathbb{E}[g(X)|X_i] - g_0 \\ g_{ij} &= g_{ij}(X_i, X_j) = \mathbb{E}[g(X)|X_i, X_j] - g_i - g_j - g_0 \\ &\ldots \\ g_{12\ldots n} &= g_{12\ldots n}(X_1, X_2, \ldots, X_n) \\ &= g(X) - g_{12\ldots n-1} - \ldots - g_i - g_j - g_0 \end{aligned}$$

where $X_{\sim i}$ denotes the vector of all input variables except X_i, and $f(\cdot)$ and $f(\cdot|\cdot)$ denote the unconditional probability density function and conditional probability density function, respectively. Then, we have:

$$\begin{aligned} \mathbb{E}[g_i(X_i)] &= \mathbb{E}[g_{ij}(X_i, X_j)] = \ldots \\ &= \mathbb{E}[g_{12\ldots n}(X_1, X_2, \ldots, X_n)] = 0. \end{aligned}$$

If the input vector $X = (X_1, X_2, \ldots, X_n)$ is independent, then we can show that

$$Cov(g_i(X_i), g_j(X_j)) = Cov(g_i(X_i), g_{ij}(X_{ij})) = 0$$

where $i \neq j$. The expansion of $g(X)$ is unique, which can lead to a unique ANOVA-HDMR decomposition of $\mathbb{V}[Y]$ as follows:

$$\mathbb{V}[Y] = \mathbb{V}[g(X)] = \sum_i \mathbb{V}[g_i(X_i)] + \sum_{j>i} \mathbb{V}[g_{ij}(X_i, X_j)] + \cdots + \mathbb{V}[g_{12\ldots n}(X_1, X_2, \ldots, X_n)].$$

Dividing both sides of the equation by $\mathbb{V}[Y]$, we obtain:

$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij} + \ldots + S_{12\ldots k} = 1. \quad (2)$$

The unique decomposition obtained by ANOVA-HDMR is the variance-based method, which is often referred to as the Sobol method or Sobol indices. The first-order effect index S_i can be rewritten as:

$$S_i = \frac{\mathbb{V}[g_i(X_i)]}{\mathbb{V}[Y]} = \frac{\mathbb{V}[Y|X_i]}{\mathbb{V}[Y]} \quad (3)$$

in which a larger $\mathbb{V}[\mathbb{E}[Y|X_i]]$ indicates a larger importance of X_i to the output variation. Apparently, the variance of the conditional expectation $\mathbb{V}[\mathbb{E}[Y|X_i]]$ is the only key item in the first-order effect index, which can be considered as a summary measure of feature importance in this case.

B. EXTENSION OF SOBOL METHOD

For the input vector $X = (X_1, X_2, \ldots, X_n)$, we denote the "standard deviation(SD)-based index" with respect to X_i by:

$$\begin{aligned} r_i^{SD} &= \sqrt{\mathbb{V}[\mathbb{E}[Y|X_i]]} = \sqrt{\mathbb{E}[\mathbb{E}[Y|X_i] - \mathbb{E}[\mathbb{E}[Y|X_i]]]^2} \\ &= \sqrt{\mathbb{E}[\mathbb{E}[Y|X_i] - \mathbb{E}[Y]]^2} \end{aligned} \quad (4)$$

in which $\mathbb{V}[\mathbb{E}[Y|X_i]]$ is also used as a summary measure in the Sobol method according to (3). In this article, r_i^{SD} is used as an index to represent the Sobol method, which is a variance-based method obtained via ANOVA-HDMR, instead of the first-order effect index S_i in (3).

In this study, the variance of the conditional expectation $\mathbb{V}[\mathbb{E}[Y|X_i]]$ is not the only method for measuring global feature importance. As mentioned before, variance is not a robust measure in robust statistics. The existence of outliers in input data will make the measurement difficult due to the poor robustness of variance. Therefore, we can use the mean absolute deviation (MAD), which is referred to as the "mean deviation" or "average absolute deviation", to address this problem.

We thus expand an SD-based index by MAD. For input vector $X = (X_1, X_2, \ldots, X_n)$, we denote the "MAD-based index" with respect to X_i as:

$$r_i^{MAD} = \mathbb{E}[|\mathbb{E}[Y|X_i] - \mathbb{E}[\mathbb{E}[Y|X_i]]|] = \mathbb{E}[|\mathbb{E}[Y|X_i] - \mathbb{E}[Y]|] \tag{5}$$

When the observation values of $\mathbb{E}[Y|X_i]$ are oscillatory, we can also use the conditional median $\mathbb{M}[Y|X_i]$ as a surrogate value in the "MAD-based index". Then we can obtain another "MAD-based index" as:

$$r_i^{MAD2} = \mathbb{E}[|\mathbb{M}[Y|X_i] - \mathbb{E}[\mathbb{M}[Y|X_i]]|]. \tag{6}$$

For a normal distribution, the sample variance of the standard deviation is below MAD [48]. However, MAD is a more robust measure for variables with contaminated distributions [49] and is widely used because MAD is easy to calculate (it avoids calculating the square value), easy to understand [50], and more tolerant of extreme values compared to standard deviation. Thus, the MAD-based index is a more robust method than the SD-based index in feature importance measurements. To our best knowledge, few researches have addressed the problem of robustness, so we compare our MAD-based method with the traditional method, the Sobol method, which does not consider robustness.

C. COMPUTATIONAL ISSUES

It should be noted especially that, in classification tasks, Y represents the prediction probability of a specific class in a given model $g(X)$ but not the real labels of class. To numerically compute the novel feature importance indices, one has to (i) obtain a sample set of inputs from input space $X = (X_1, X_2, \ldots, X_n)$ and calculate the corresponding output by $Y = g(X)$, (ii) compute the conditional mean or median of Y with respect to $X_1, \ldots, X_n$, separately, with the samples in step (i), and (iii) compute the MAD-based indices r_i^{MAD}, r_i^{MAD2} and the SD-based index r_i^{SD} using the results from step (ii).

In step (ii), we obtain the conditional values using the Bins method [6], which allows us to calculate both continuous and discrete inputs. We use the output and input samples $(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{N\times(n+1)}$ in step (i) to calculate the conditional mean value $\mathbb{E}[Y|X_i]$ or conditional median value $\mathbb{M}[Y|X_i]$. When X_i is a continuous variable, we can easily obtain the scatterplot of X_i and Y. Next, we can partition the X_i-axis of the scatterplot on the horizontal plane. We then divide X_i into M mutually exclusive subsets $\mathcal{X}_i^m (m = 1, 2, \ldots, M)$, where $\cup_{m=1}^M \mathcal{X}_i^m = X_i$, $\mathcal{X}_i^m \cap \mathcal{X}_i^q = \emptyset, (m \neq q)$. Then, we substitute the point condition value $X_i = x_i$ with the bin condition value $X_i \in \mathcal{X}_i^m$. Formally, we use the bin conditional value $\mathbb{E}[Y|X_i \in \mathcal{X}_i^m]$ or $\mathbb{M}[Y|X_i \in \mathcal{X}_i^m]$ to replace the point conditional value $\mathbb{E}[Y|X_i = x_i]$ or $\mathbb{M}[Y|X_i = x_i]$. In addition, we can obtain the bins estimator as follow:

$$\mathbb{E}[Y|X_i \in \mathcal{X}_i^m] = \frac{1}{N_m} \sum_{x\in\mathcal{X}_i^m} \mathcal{Y}_x$$

$$\mathbb{M}[Y|X_i \in \mathcal{X}_i^m] = \underset{x\in\mathcal{X}_i^m}{median}\, \mathcal{Y}_x$$

where N_m indicates the number of observations falling in the m^{th} interval set $\mathcal{X}_i^m$, and $\mathcal{Y}_x$ denotes the output $\mathcal{Y}$ corresponding to one observation $x \in \mathcal{X}_i^m$. However, if the input X_i is a discrete variable, then we can only divide levels or values of X_i into M' mutually exclusive subsets $\mathcal{X}_i^m (m = 1, 2, \ldots, M')$, where M' is the number of the different levels or values in X_i. The other calculation manipulation is the same as the continuous inputs in X.

In the last step, with the conditional mean value $\mathbb{E}[Y|X_i]$ or conditional median value $\mathbb{M}[Y|X_i]$ calculated in step (ii), we can calculate all indices in Section III-B by

$$r_i^{SD} = \sqrt{\sum\nolimits_{i=1}^{M} \frac{N_m}{N} (\mathbb{E}[Y|X_i \in \mathcal{X}_i^m] - \mathbb{E}[Y])^2} \tag{7}$$

$$r_i^{MAD} = \sum\nolimits_{i=1}^{M} \frac{N_m}{N} |\mathbb{E}[Y|X_i \in \mathcal{X}_i^m] - \mathbb{E}[Y]| \tag{8}$$

$$\begin{aligned} r_i^{MAD2} = {} & \sum_{i=1}^{M} \frac{N_m}{N} |\mathbb{M}[Y|X_i \in \mathcal{X}_i^m] \\ & - \sum_{m=1}^{M} \frac{N_m}{N} \mathbb{M}[Y|X_i \in \mathcal{X}_i^m]| \end{aligned} \tag{9}$$

where $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X_i]] = \sum_{m=1}^{M} \frac{N_m}{N} \mathbb{E}[Y|X_i \in \mathcal{X}_i^m] = \sum_{i=1}^{N} \mathcal{Y}_i$.

IV. EXPERIMENTS

In this section, the robustness of the proposed MAD-based indices is shown in two experiments: 1) artificial models with randomly generated datasets and 2) machine learning models built on public datasets. These experiments allow us to calculate and contrast feature importance so that we can conveniently test the robustness of the proposed methods. For all models in this section, the feature importance derived by either the MAD- or SD-based index is compared.

A. FEATURES WITH CONTAMINATED DISTRIBUTION

To evaluate the robustness of the proposed method on data with outliers, we generated data with contaminated distributions in the experiments with artificial models or added contaminated data to real data in the experiments with machine learning models. The contaminated distribution of a feature, X_i, is defined by its original distribution $F(X_i)$ and an error distribution $N(\theta, \sigma^2)$

$$\begin{aligned} P(X_i < x) &= F^{mixed}(\mu, \sigma_0^2, \sigma^2, \epsilon) \\ &= (1-\epsilon)F(x) + \epsilon\Phi(\frac{x-\mu}{\sigma}) \end{aligned} \tag{10}$$

where μ and σ_0^2 denote the mean and variance of X_i, respectively; $\Phi(\cdot)$ denotes the cumulative probability density function of a standard normal distribution; $\epsilon \in [0, 1]$ indicates the percentage of samples X_i generated by the error distribution $N(\mu, \sigma^2)$. The variance σ^2 is often set to be much larger than σ_0^2, and the parameter ϵ is typically set to below 0.05 when we assume that a feature with contaminated distribution has tiny errors.

B. ARTIFICIAL MODELS

In this subsection, the efficiency of the proposed indices is shown by two models

$$f_1(X) = X_1^2 + X_2^2,$$
$$f_2(X) = X_1 + X_2 + X_1X_2.$$

These simple models allow us to easily calculate and contrast the feature importance and error rates of inputs. Then, we can evaluate the robustness of the proposed indices. For both models, we first, assume that X_1 and X_2 follow the normal distributions $N(1, 1)$ and $N(1, 1.2^2)$, respectively. Thus, we can obtain the theoretical values of the SD-based index of $X_i(i = 1, 2)$ for both models:

$$r_i^{the} = \begin{cases} \sqrt{V(X_i^2)} = \sqrt{2\sigma_0^4 + 4\mu^2\sigma_0^2}, & for\ f_1(X) \\ |1 + E(X_j)|\sqrt{V(X_i)} = |1 + \mu|\sigma_0, & i \neq j,\ for\ f_2(X) \end{cases} \tag{11}$$

where μ and σ_0^2 denote the mean and variance of X_i, respectively.

Second, to test the robustness of the proposed MAD-based index, we assume that X_1 has tiny errors in its distribution and is generated by the contaminated distribution (refer to (10)):

$$\begin{aligned} P(X_1 < x) &= F^{mixed}(\mu = 1, \sigma_0^2 = 1, \sigma^2, \epsilon) \\ &= (1 - \epsilon)\Phi(\frac{x-1}{1}) + \epsilon\Phi(\frac{x-1}{\sigma}) \end{aligned} \tag{12}$$

For each feature, 1,000 values are generated, and then the MAD-based indices, r_i^{MAD} and r_i^{MAD2}, and the SD-based index, r_i^{SD}, are calculated. The process is repeated 300 times to evaluate the error rate or confidence interval of feature importance for each index.

1) ANALYZING ROBUSTNESS OF ALL INDICES BY ERROR RATES

When no errors occurred in the distribution, $X_1 \sim N(1, 1^2)$ and $X_2 \sim N(1, 1.2^2)$. For each feature, 1,000 values are generated. Because X_1 and X_2 are equivalent in both models, and the variance of X_2 is above that of X_1, the theoretical feature importance of X_2 is above that of X_1 for both models according to (11). After tiny errors are added to X_1 based on the contaminated distribution defined in (12), the r_i^{MAD}, r_i^{MAD2} and r_i^{SD} of X_1 and X_2 are calculated based on the data using the estimation method described in Section III-C. The error rate of the SD-based index is defined as:

$$Err^{SD} = \frac{1}{300}\sum_{k=1}^{300} \mathbb{I}(r_1^{SD}(k) > r_2^{SD}(k)) \tag{13}$$

where $r_i^{SD}(k)$ denotes the SD-based index of X_i in the k^{th} experiment, and $\mathbb{I}(*)$ denotes the indication function. Similarly, the error rate for the MAD-based index can be calculated. The error rates can be used to evaluate the robustness of the feature importance indices. A low error rate indicates that the index is robust to errors in the contaminated distribution.

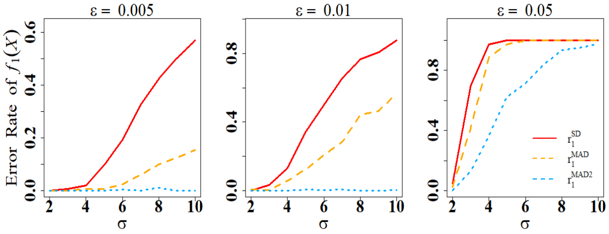


FIGURE 1. Error rates of $f_1(X)$ by different indices.

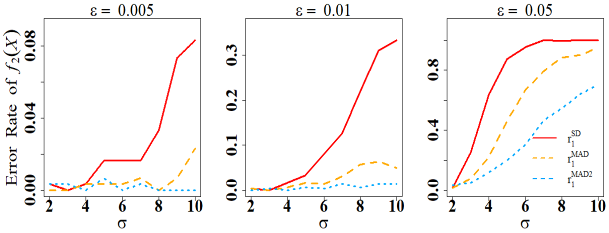


FIGURE 2. Error rates of $f_2(X)$ by different indices.

The results of the error rates under different values of σ and ϵ are shown in Figures 1 and 2. When the standard error σ increases to 10, the error rates of all indices are increasingly closer to 1. As the resultant curves show that the error rates of the MAD-based indices are below that of the SD-based index, the MAD-based indices are more robust than the SD-based index for both models. The index r_i^{MAD2} is shown to be the most robust index in the estimation of the error rates. Also, when the contamination rate ϵ increases, the difference in error rates among the three indices becomes negligible.

2) ANALYZING ROBUSTNESS OF ALL INDICES BY FEATURE IMPORTANCE

After computing a quantitative contrast of the error rates, we compare the magnitude of feature importance of X_1 generated by all indices. In the original distribution, the theoretical values of the SD-based index of X_1 for $f_1(X)$ and $f_2(X)$ are $\sqrt{6}$ and 2 (refer to (11)), respectively. After the errors are added into X_1, the MAD-based indices must be adjusted for comparison with the theoretical values. Because the sample mean of the mean absolute deviation for X_1 is $\sqrt{(n-1)/n}\sqrt{2/\pi}\sigma_0$ [48], the MAD-based indices are lower than expected. Thus, we multiply the MAD-based indices by a constant $\sqrt{2/\pi}$ to distinguish the effect of the magnitude of feature importance.

With the generated input samples ($X_1 \sim F^{mixed}(\theta = 1, \sigma_0^2 = 1, \sigma^2, \epsilon), X_2 \sim N(1, 1.2^2)$), the estimates of the feature importance for both models are reported in Figures 3 and 4. As σ approaches 10, all indices grow rapidly, and the 90% confidence interval becomes large. The resultant curves show that the feature importance of the MAD-based indices are blow those of the SD-based index; thus, the MAD-based indices are more robust than the SD-based index.

C. MACHINE LEARNING MODELS

We begin this subsection by contrasting the efficiency of the proposed indices to the other indices when using complex machine learning models, including supported vector

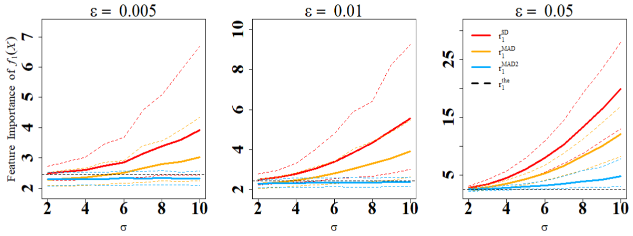


FIGURE 3. Feature importance values of $f_1(X)$ by different indices.

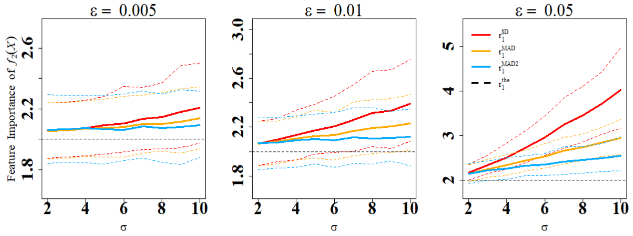


FIGURE 4. Feature importance values of $f_2(X)$ by different indices.

TABLE 1. Description of the Dimaonds data.

Feature	Description
price	Price in US dollars.
carat	Weight of the diamond.
cut	Quality of the cut.
color	Diamond colour.
clarity	A measurement of how clear the diamond is.
x	Length in mm.
y	Width in mm.
z	Depth in mm.
depth	Total depth percentage = z / mean(x, y).
table	Width of the top of the diamond relative to the widest point.

machine (SVM), neural network (NN) and xgboost (XGB). These complex models allow us to evaluate the robustness of the proposed indices.

1) DATASETS AND MACHINE LEARNING MODELS

We train the machine learning models on two real-world datasets: the first contains data about Diamonds,[1] including 54,000 records with the output target of *price* and nine input features (refer to Table 1); and the second contains data about Boston housing,[2] including 506 entries with the output target of *medv* and 13 input features for homes from various suburbs in Boston (refer to Table 2). To improve the generalization, parameters are tuned during training for both data. During the training of machine learning models, 5-fold cross-validation is used for model evaluation. The NN is trained as a neural network with 3 layers and 5 hidden neurons, and its activation function is sigmoid. The SVM uses the default parameters set in R for regression task with C equals 1. And the XGB also uses the default parameters set in R for the linear booster with maximum number of iterations equals 150. All categorical features are transformed into numerical values and normalized via regression.

TABLE 2. Description of the Boston housing data.

Feature	Description
medv	Median price of owner-occupied homes.
crim	Per capita crime rate by town.
zn	Proportion of residential land zoned
indus	Proportion of non-retail business acres per town.
chas	Charles River dummy variable.
nox	Nitrogen oxides concentration.
rm	Average number of rooms per dwelling.
age	Proportion of owner-occupied units built prior to 1940.
dis	Weighted mean of distances to five Boston employment centers.
rad	Index of accessibility to radial highways.
tax	Full-value property-tax rate.
ptratio	Pupil-teacher ratio by town.
black	The proportion of blacks by town.
lstat	Lower status of the population.

After the models are trained, we calculate the output values of Y based on the trained models with X. To test the robustness of the proposed MAD-based index, we assume that an arbitrary feature $X_i(X_i \in X)$ has tiny errors in its original distribution and is generated by a contaminated distribution (refer to (10)):

$$\begin{aligned} P(X_i < x) &= F^{mixed}\left(\hat{\mu}_i, \hat{\sigma}_i^2, \sigma^2 = (5\hat{\sigma}_i)^2, \epsilon = 0.05\right) \\ &= (1-\epsilon)\, F(x) + \epsilon \Phi\left(\frac{x-\hat{\mu}_i}{\sigma}\right) \end{aligned} \tag{14}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ represent the sample mean and variance of X_i, respectively. After X_i is contaminated, we obtain a new data set $\hat{X}^i$ and calculate the corresponding output $\hat{Y}^i$ based on the trained model.

2) ANALYZING THE ROBUSTNESS OF ALL INDICES ON MACHINE LEARNING MODELS

With the Diamonds data, we specify the first feature "carat" of diamonds as the contaminated feature to contrast the robustness of the proposed indices on the trained machine learning models. We then generate the contaminated data $\hat{X}^{carat}$ and the output $\hat{Y}^{carat}$ according to (14). The r_i^{MAD}, r_i^{MAD2} and r_i^{SD} of each feature are calculated based on the contaminated data $(\hat{X}^{carat}, \hat{Y}^{carat})$ and the original data, respectively, using the estimation method described in Section III-C. The percent variation of the SD-based index for the feature "carat" is defined as r_i^{SD} of the contaminated data divided by that of the original data. Similarly, the percent variation of the MAD-based index can be calculated, and then can be used to evaluate the robustness of the MAD-based method compared with the percent variation of SD-based method.

First, the percent variation of the feature importance for the feature "carat" are shown in Table 3. The resultant table shows that the variation percentages of the MAD-based indices in the feature "carat" are markedly below those of the SD-based index. Apparently, the MAD-based indices are more robust than the SD-based index in the estimation of the percent variation of the feature "carat". When we specify the feature "carat" as the contaminated feature, the average variation percentages of the MAD-based indices are also below that of the SD-based index, while r_i^{MAD2} is much

[1] https://www.kaggle.com/shivam2503/diamonds

[2] https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

TABLE 3. Percent variation in feature importance values with contaminated data $\hat{X}^{carat}$.

Models	Index	Input features									Average
		carat	cut	color	clarity	depth	table	x	y	z	
NN	SD	**30.9%**	20.2%	23.2%	27.0%	24.1%	18.0%	12.3%	12.3%	12.4%	**20.1%**
	MAD	**15.9%**	21.5%	23.6%	31.7%	21.2%	19.7%	15.8%	15.8%	16.0%	**20.0%**
	MAD2	**16.8%**	3.8%	5.4%	5.2%	4.5%	4.0%	2.2%	2.2%	2.3%	**5.1%**
SVM	SD	**91.7%**	43.0%	40.8%	45.6%	48.7%	37.8%	32.9%	32.8%	33.1%	**45.2%**
	MAD	**34.6%**	43.2%	40.0%	48.1%	39.1%	40.7%	34.5%	34.4%	34.9%	**38.8%**
	MAD2	**31.1%**	2.9%	5.4%	4.3%	3.6%	4.5%	1.3%	1.3%	1.4%	**6.2%**
XGB	SD	**40.4%**	26.6%	26.2%	31.1%	27.3%	22.5%	18.0%	18.0%	18.1%	**25.3%**
	MAD	**18.1%**	23.3%	26.4%	28.1%	22.0%	25.1%	18.1%	18.0%	18.1%	**21.9%**
	MAD2	**17.2%**	5.5%	4.6%	6.0%	4.9%	4.3%	2.0%	1.9%	2.0%	**5.4%**

TABLE 4. Percent variations in feature importance values with contaminated data of another 3 features with most feature importance values with the Diamonds dataset.

Models	Index	$\hat{X}^{color}$		$\hat{X}^{table}$		$\hat{X}^{x}$	
		color	Average	table	Average	x	Average
NN	SD	**54.4%**	9.2%	**103.9%**	17.1%	**0.3%**	14.8%
	MAD	**19.2%**	4.9%	**37.6%**	9.6%	**0.4%**	15.6%
	MAD2	**8.3%**	2.6%	**19.1%**	5.2%	**2.9%**	4.9%
SVM	SD	**0.2%**	0.5%	**2.9%**	1.8%	**6.1%**	5.1%
	MAD	**0.1%**	0.6%	**4.0%**	2.0%	**4.1%**	5.1%
	MAD2	**0.5%**	0.7%	**3.6%**	1.6%	**4.1%**	6.2%
XGB	SD	**7.2%**	1.5%	**35.1%**	6.6%	**1.5%**	3.2%
	MAD	**5.8%**	1.5%	**19.6%**	4.8%	**2.7%**	3.2%
	MAD2	**3.3%**	1.5%	**13.3%**	3.0%	**2.8%**	2.3%

TABLE 5. Percent variation in feature importance values with contaminated data $\hat{X}^{crim}$.

Models	Index	Input features													Average
		crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	
NN	SD	**77.3%**	16.5%	9.7%	3.6%	3.7%	5.4%	8.0%	9.2%	17.1%	5.7%	10.7%	1.1%	4.8%	**13.3%**
	MAD	**35.7%**	20.7%	7.0%	3.6%	6.3%	9.3%	11.6%	8.0%	21.1%	0.5%	7.1%	4.6%	9.1%	**11.1%**
	MAD2	**18.4%**	0.1%	0.2%	5.5%	0.6%	1.7%	1.2%	1.3%	2.7%	0.7%	0.3%	8.2%	0.9%	**3.2%**
SVM	SD	**14.3%**	2.3%	3.5%	2.6%	5.3%	3.7%	5.8%	9.8%	7.5%	4.3%	3.9%	11.0%	2.9%	**5.9%**
	MAD	**5.4%**	3.4%	4.2%	2.6%	3.4%	2.9%	5.7%	5.1%	5.6%	5.3%	5.4%	5.1%	3.3%	**4.4%**
	MAD2	**2.8%**	0.2%	1.1%	6.6%	1.0%	0.6%	1.3%	0.5%	4.0%	0.4%	0.8%	1.6%	0.0%	**1.6%**
XGB	SD	**6.2%**	1.3%	1.7%	0.6%	2.0%	1.6%	2.5%	3.2%	3.7%	2.1%	1.8%	3.2%	1.6%	**2.4%**
	MAD	**5.7%**	1.6%	1.9%	0.6%	2.0%	1.3%	2.2%	2.5%	3.2%	2.5%	2.4%	2.5%	1.6%	**2.3%**
	MAD2	**0.4%**	0.0%	0.3%	13.0%	2.6%	1.2%	1.7%	1.0%	0.9%	1.6%	0.9%	1.1%	0.3%	**1.9%**

below r_i^{SD}. This result shows that in the estimation of the average precent variation, MAD-based indices are still more robust than the SD-based index, while the efficiency of the MAD-based index r_i^{MAD2} in robustness is highest.

Second, for comparison, we specify another three features with the most feature important values as the contaminated feature, separately, and show its corresponding percent variation of the contaminated feature in Table 4. For simplicity, with respect to each contaminated data of specified feature, table 4 only illustrate its variation percentages in contaminated feature and the average variation percentages. In the estimation of the percent variations of all three chosen features, the MAD-based indices yield better robustness than the SD-based index for the percent variations of the SD-based index that are higher than 10%. In the estimation of the average percent variations, MAD-based indices are still markedly more robust than the SD-based index for other cases, even though the performance of MAD-based indices is marginally unstable in robustness compared to that of the SD-based index when the percent variations are below 10%.

With the Boston housing data, we also specify the first feature "crim" as the contaminated feature and generate the contaminated data X^{crim} and the output Y^{crim} based on (14), as with the diamond data. The percent variations of the feature importance for the feature "crim" are shown in Table 5. We also choose another three features with the most feature-important values as the contaminated feature, separately, and show its corresponding percent variation of the contaminated feature in Table 6. The results of the Boston data are nearly identical to those of the diamond data when comparing the robustness between the MAD-based indices and SD-based index. The results of the MAD-based indices are more robust than those of the SD-based index for nearly all cases shown in Tables 5 and 6.

The experimental results on both the Diamonds and Boston housing data demonstrate the advantages of the proposed MAD-based method. First, we can find the efficiency of the proposed MAD-based method in robustness by comparing the results of MAD-based indices with that of the SD-based index, which is an index of the Sobol method.

TABLE 6. **Percent variations in feature importance values with contaminated data of another 3 features with most feature importance values with the Boston housing dataset.**

Models	Index	$\hat{X}^{indus}$		$\hat{X}^{rm}$		$\hat{X}^{lstat}$	
		indus	**Average**	**rm**	**Average**	**lstat**	**Average**
NN	SD	**2.55%**	1.59%	**23.57%**	13.14%	**75.67%**	27.23%
	MAD	**0.54%**	0.99%	**11.36%**	10.70%	**23.77%**	22.74%
	MAD2	**2.31%**	1.20%	**11.68%**	2.75%	**30.24%**	6.80%
SVM	SD	**16.90%**	10.33%	**66.13%**	28.69%	**10.82%**	7.27%
	MAD	**10.12%**	8.35%	**22.81%**	19.53%	**8.74%**	5.39%
	MAD2	**7.11%**	2.81%	**13.85%**	2.47%	**7.69%**	1.87%
XGB	SD	**2.49%**	0.28%	**12.37%**	7.49%	**18.06%**	10.39%
	MAD	**0.97%**	0.16%	**6.90%**	6.12%	**12.11%**	9.95%
	MAD2	**0.05%**	0.15%	**8.86%**	1.63%	**13.60%**	3.82%

For example, in the estimation of the percent variation of a specified feature, MAD-based indices are markedly more efficient in robustness for variation percentages greater than 10%. Second, we can also use the MAD-based method as a reliable and model-agnostic method to estimate the feature importance values of complex models instead of the Sobol method by implementing the two indices of the MAD-based method. For example, in the estimation of both the percent variations and the average percent variations, MAD-based indices are only marginally unstable in robustness compared to SD-based index when the percent variations are below 10%, while MAD-based indices are markedly more efficient in robustness in other cases.

V. CONCLUSION

In this article, we proposed a model-agnostic method to measure global feature importance. The method is based on high-dimensional model representation and is an extension of the standard deviation-based method, which is a variance-based method obtained via ANOVA-HDMR. The proposed method is more robust than the SD-based method.

The proposed method can be used in various domains because it is generally applicable to complex models and can provide similar explanations when inputs have unanticipated small modifications or contamination. The main limitation of this study is that the method assumes feature independence and thus cannot be used to estimate correlated features. Thus, future work should investigate a method to estimate feature importance when features are assumed to be correlated.

REFERENCES

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Jan. 2019, doi: 10.1145/3236009.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[3] E. Štrumbelj and I. Kononenko, "A general method for visualizing and explaining black-box regression models," in *Proc. Int. Conf. Adapt. Natural Comput. Algorithms*. Berlin, Germany: Springer, 2011.

[4] Z. C. Lipton, "The mythos of model interpretability," (in English), *Commun. ACM*, vol. 61, no. 10, pp. 36–43, Sep. 2018, doi: 10.1145/3233231.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*. [Online]. Available: http://arxiv.org/abs/1702.08608

[6] E. Borgonovo and E. Plischke, "Sensitivity analysis: A review of recent advances," *Eur. J. Oper. Res.*, vol. 248, no. 3, pp. 869–887, Feb. 2016, doi: 10.1016/j.ejor.2015.06.032.

[7] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2018.

[8] T. Aven, "Risk assessment and risk management: Review of recent advances on their foundation," (in English), *Eur. J. Oper. Res.*, vol. 253, no. 1, pp. 1–13, Aug. 2016, doi: 10.1016/j.ejor.2015.12.023.

[9] G. Li and H. Rabitz, "General formulation of HDMR component functions with independent and correlated variables," *J. Math. Chem.*, vol. 50, no. 1, pp. 99–130, Jan. 2012, doi: 10.1007/s10910-011-9898-0.

[10] T. A. Mara and S. Tarantola, "Variance-based sensitivity indices for models with dependent inputs," (in English), *Rel. Eng. Syst. Saf.*, vol. 107, pp. 115–121, Nov. 2012, doi: 10.1016/j.ress.2011.08.008.

[11] L. Breiman, "Random forests," (in English), *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[12] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statist. Comput.*, vol. 27, no. 3, pp. 659–678, May 2017, doi: 10.1007/s11222-016-9646-1.

[13] S. H. Welling, H. H. F. Refsgaard, P. B. Brockhoff, and L. H. Clemmensen, "Forest floor visualizations of random forests," 2016, *arXiv:1605.09196*. [Online]. Available: http://arxiv.org/abs/1605.09196

[14] O. F. Alis and H. Rabitz, "Efficient implementation of high dimensional model representations," (in English), *J. Math. Chem.*, vol. 29, no. 2, pp. 127–142, 2001, doi: 10.1023/A:1010979129659.

[15] I. M. Sobol, "Theorems and examples on high dimensional model representation," (in English), *Rel. Eng. Syst. Saf.*, vol. 79, no. 2, pp. 187–193, Feb. 2003, doi: 10.1016/S0951-8320(02)00229-6.

[16] E. Borgonovo, "Sensitivity analysis with finite changes: An application to modified EOQ models," (in English), *Eur. J. Oper. Res.*, vol. 200, no. 1, pp. 127–138, Jan. 2010, doi: 10.1016/j.ejor.2008.12.025.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier" in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[18] W. Duivesteijn and J. Thaele, "Understanding where your classifier does (Not) work–the SCaPE model class for EMM," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 809–814.

[19] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions predict," presented at the ICLR, 2017.

[20] S. Avanti, G. Peyton, and K. Anshul, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," presented at the ICML Workshop Hum. Interpretability Mach. Learn., 2016.

[22] H. Xiao and Y. Duan, "Sensitivity analysis of correlated inputs: Application to a riveting process model," *Appl. Math. Model.*, vol. 40, nos. 13–14, pp. 6622–6638, Jul. 2016, doi: 10.1016/j.apm.2016.02.008.

[23] M. Možina *et al.*, "Nomograms for visualization of naive Bayesian classifier," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Berlin, Germany: Springer, 2004.

[24] F. Poulet, "SVM and graphical algorithms: A cooperative approach," (in English), in *Proc. 4th IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2004, pp. 499–502, doi: 10.1109/ICDM.2004.10068.

[25] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140, doi: 10.1371/journal.pone.0130140.

[26] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," 2017, *arXiv:1711.06104*. [Online]. Available: http://arxiv.org/abs/1711.06104

[27] B. Ghazi, R. Panigrahy, and J. R. Wang, "Recursive sketches for modular deep learning," presented at the 36th Int. Conf. Mach. Learn., 2019.

[28] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," (in English), *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.

[29] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," (in English), *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Dec. 2014, doi: 10.1007/s10115-013-0679-x.

[30] G. Li, C. Rosenthal, and H. Rabitz, "High dimensional model representations," *J. Phys. Chem. A*, vol. 105, no. 33, pp. 7765–7777, Aug. 2001.

[31] G. Y. Li, S. W. Wang, H. Rabitz, S. Y. Wang, and P. Jaffe, "Global uncertainty assessments by high dimensional model representations (HDMR)," (in English), *Chem. Eng. Sci.*, vol. 57, no. 21, pp. 4445–4460, Nov. 2002, doi: 10.1016/S0009-2509(02)00417-7.

[32] R. L. Iman and S. C. Hora, "A robust measure of uncertainty importance for use in fault tree system analysis," *Risk Anal.*, vol. 10, no. 3, pp. 401–406, Sep. 1990.

[33] A. Jakulin, M. Možina, J. Demšar, I. Bratko, and B. Zupan, "Nomograms for visualizing support vector machines," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2005, pp. 108–117.

[34] M. Gevrey, I. Dimopoulos, and S. Lek, "Review and comparison of methods to study the contribution of variables in artificial neural network models," *Ecol. Model.*, vol. 160, no. 3, pp. 249–264, Feb. 2003.

[35] F. Recknagel, M. French, P. Harkonen, and K. I. Yabunaka, "Artificial neural network approach for modelling and prediction of algal Blooms," *Ecol. Model.*, vol. 96, nos. 1–3, pp. 11–28, 1997.

[36] J. C. Helton, "Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal," *Rel. Eng. Syst. Saf.*, vol. 42, nos. 2–3, pp. 327–367, Jan. 1993.

[37] A. Saltelli and J. Marivoet, "Non-parametric statistics in sensitivity analysis for model output: A comparison of selected techniques," *Rel. Eng. Syst. Saf.*, vol. 28, no. 2, pp. 229–253, Jan. 1990.

[38] J. C. Helton and F. J. Davis, "Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems," *Rel. Eng. Syst. Saf.*, vol. 81, no. 1, pp. 23–69, Jul. 2003.

[39] H. Rabitz and O. F. Alis, "General foundations of high-dimensional model representations," (in English), *J. Math. Chem.*, vol. 25, nos. 2–3, pp. 197–233, 1999, doi: 10.1023/A:1019188517934.

[40] G. Y. Li, S. W. Wang, C. Rosenthal, and H. Rabitz, "High dimensional model representations generated from low dimensional data samples. 1. mp-cut-HDMR," (in English), *J. Math. Chem.*, vol. 30, no. 1, pp. 1–30, Jul. 2001, doi: 10.1023/A:1013172329778.

[41] G. Chastaing, F. Gamboa, and C. Prieur, "Generalized Hoeffding-Sobol decomposition for dependent variables–application to sensitivity analysis," (in English), *Electron. J. Statist.*, vol. 6, pp. 2420–2448, 2012, doi: 10.1214/12-EJS749.

[42] S. Kucherenko, S. Tarantola, and P. Annoni, "Estimation of global sensitivity indices for models with dependent variables," (in English), *Comput. Phys. Commun.*, vol. 183, no. 4, pp. 937–946, Apr. 2012, doi: 10.1016/j.cpc.2011.12.020.

[43] E. Borgonovo, "A new uncertainty importance measure," *Rel. Eng. Syst. Saf.*, vol. 92, no. 6, pp. 771–784, Jun. 2007.

[44] G. C. Critchfield and K. E. Willard, "Probabilistic analysis of decision trees using Monte Carlo simulation," *Med. Decis. Making*, vol. 6, no. 2, pp. 85–92, Jun. 1986.

[45] P. Young, "Data-based mechanistic modelling, generalised sensitivity and dominant mode analysis," *Comput. Phys. Commun.*, vol. 117, nos. 1–2, pp. 113–129, Mar. 1999.

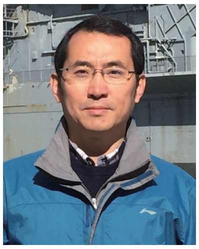
[46] P. J. Huber, *Robust Statistical Procedures*. Philadelphia, PA, USA: SIAM, 1996.

[47] I. M. Sobol, "Sensitivity estimates for nonlinear mathematical models," *Math. Model. Comput. Exp.*, vol. 1, no. 4, pp. 407–414, 1990.

[48] R. A. Fisher, "A mathematical examination of the methods of determining the accuracy of observation by the mean error, and by the mean square error," *Monthly Notices Roy. Astronomical Soc.*, vol. 80, no. 8, pp. 758–770, Jun. 1920.

[49] J. W. Tukey, "A survey of sampling from contaminated distributions," in *Contributions to Probability and Statistics*. Stanford, CA, USA: Stanford Univ. Press, 1960, pp. 448–485.

[50] S. Nahmias and T. L. Olsen, *Production and Operations Analysis*. Long Grove, IL, USA: Waveland Press, 2015.



XIAOHANG ZHANG (Member, IEEE) received the Ph.D. degree in management science and engineering from the Beijing University of Posts and Telecommunications in 2003. From 2015 to 2017, he was a Research Fellow with the Department of Statistics, University of Michigan at Ann Arbor. He is currently a Professor in the field of information systems with the Beijing University of Posts and Telecommunications. His recent research interests include machine learning, business intelligence, and risk modeling.

LING WU received the B.Admin. and master's degrees from the Beijing University of Posts and Telecommunications in 2010 and 2013, respectively, where he is currently pursuing the Ph.D. degree with the School of Economics and Management. His research interests include machine learning and business intelligence.

ZHENGREN LI received the Ph.D. degree in management science and engineering from the Beijing University of Posts and Telecommunications (BUPT) in 2014. He is currently an Associate Professor with BUPT. His research interests include user behavior analysis, data mining, and business intelligence.

HUAYUAN LIU was born in 1989. He received the B.S. and M.S. degrees from Beihang University, in 2010 and 2013, respectively. He is currently a Senior Engineer with the China North Vehicle Research Institute. His research interests include behavior analysis and machine learning.