Structural and Multidisciplinary Optimization (2019) 59:1439–1454
https://doi.org/10.1007/s00158-018-2137-6

RESEARCH PAPER

Variable selection using Gaussian process regression-based metrics for high-dimensional model approximation with limited data

Kyungeun Lee[1] · Hyunkyoo Cho[2] · Ikjin Lee[1]

Received: 20 June 2018 / Revised: 23 October 2018 / Accepted: 25 October 2018 / Published online: 28 November 2018

Abstract

In recent years, the importance of computationally efficient surrogate models has been emphasized as the use of high-fidelity simulation models increases. However, high-dimensional models require a lot of samples for surrogate modeling. To reduce the computational burden in the surrogate modeling, we propose an integrated algorithm that incorporates accurate variable selection and surrogate modeling. One of the main strengths of the proposed method is that it requires less number of samples compared with conventional surrogate modeling methods by excluding dispensable variables while maintaining model accuracy. In the proposed method, the importance of selected variables is evaluated using the quality of the model approximated with the selected variables only. Nonparametric probabilistic regression is adopted as the modeling method to deal with inaccuracy caused by using selected variables during modeling. In particular, Gaussian process regression (GPR) is utilized for the modeling because it is suitable for exploiting its model performance indices in the variable selection criterion. Outstanding variables that result in distinctly superior model performance are finally selected as essential variables. The proposed algorithm utilizes a conservative selection criterion and appropriate sequential sampling to prevent incorrect variable selection and sample overuse. Performance of the proposed algorithm is verified with two test problems with challenging properties such as high dimension, nonlinearity, and the existence of interaction terms. A numerical study shows that the proposed algorithm is more effective as the fraction of dispensable variables is high.

Keywords Surrogate model · Variable selection · High-dimensional problem · Gaussian process regression · Limited data

Abbreviations

n	Dimension of input
$\mathbf{X}$	Training input
$\mathbf{X}_*$	New input
$\mathbf{f}_*$	Posterior output with zero mean function
$\overline{\mathbf{f}}_*$	Best estimation for $\mathbf{f}_*$
$\mathbf{g}_*$	Posterior output with explicit basis function
$\mathrm{cov}(\mathbf{g}_*)$	Covariance of $\mathbf{g}_*$
$\mathbf{y}$	Training output (noisy response)
$m_i(x_i)$	Mean function of GPR in x_i-y plane
$c(\mathbf{x}\|\mathbf{X})$	Posterior variance in a specified point $\mathbf{x}$
m	Number of observations
$h(\mathbf{x})$	Basis function of GPR
$k(\mathbf{x})$	Covariance function of GPR
$\mathrm{cov}(\mathbf{f}_*)$	Covariance of $\mathbf{f}_*$
$\overline{\mathbf{g}}_*$	Best estimation for $\mathbf{g}_*$
$\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$	Coefficients of basis function and their estimation
$\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}$	Hyperparameters of covariance function and their estimation
$\sigma^2, \hat{\sigma}^2$	Noise variance and its estimation
$k_i(x_i, x';\boldsymbol{\theta})$	Covariance function of GPR with x_i-y plane and hyperparameter $\boldsymbol{\theta}$
$\boldsymbol{\varepsilon}$	Gaussian noise

Responsible Editor: YoonYoung Kim

✉ Ikjin Lee
ikjin.lee@kaist.ac.kr

Kyungeun Lee
angelbi@kaist.ac.kr

Hyunkyoo Cho
hyunkcho@mokpo.ac.kr

[1] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

[2] Department of Mechanical Engineering, Mokpo National University, Muan-gun 58554, South Korea

1 Introduction

In these days, performances of engineering designs are evaluated using computer simulations instead of physical testing to save cost and time for design development. As the simulation models become more complex, their computational cost has soared significantly. For the reason, surrogate modeling has been utilized to replace the time-consuming simulation models. However, for high-dimensional problems, building a surrogate model itself is a computationally expensive task since the number of samples for surrogate modeling highly depends on the dimensionality (Jin et al. 2002). The dimensionality issue can be effectively relieved by selecting and using essential variables that have great influence on the output response when we create surrogate models (Shan and Wang 2010). Therefore, for high-dimensional problems, variable selection is necessary to create an accurate surrogate model with given computational resources.

Many variable selection methods use samples produced by experimental design methods to identify the degree of relationship between variables and the output responses. More accurate and efficient variable selection is possible when the sample has high quality, i.e., orthogonality and space-filling property. To find effective sample location, experimental design has been extensively studied: input-domain-based criterion (Jin et al. 2003; Joseph et al. 2015; Joseph and Hung 2008; Pronzato and Walter 1988; Stein 1987), information-based criterion (Beck and Guillas 2016; Ko et al. 1995), and uncertainty-based criterion of posterior variance of the Gaussian process regression (GPR) model (Gorodetsky and Marzouk 2016). Some variable selection methods are paired with a specific experimental design to maximize their performance. For example, analysis of variance (ANOVA) that is the most classical variable selection method is mostly utilized in combination with the design of experiment (DOE) (Hayter 2012). Moon et al. introduced a two-stage variable selection method in combination with their own sampling technique based on Gram-Schmidt orthogonalization (Moon et al. 2012). Once the sample location is determined, output response at each sample is computed, and relevance between variables and output responses is measured. As a measure of the relevance, influence diagnostic scores such as information gain, Akaike information criterion (AIC), Bayesian information criterion (BIC), and linear/nonlinear coefficients (Qi and Zhang 2001; Helton et al. 2006) have been used. In addition, Sobol' indices, which is a global sensitivity analysis method, was developed to measure the influence of mutually interacting input variables on the output in highly nonlinear problems (Homma and Saltelli 1996; Sobol 2001). A better index compared with Sobol' indices when applied in strong interaction properties is also developed (Saltelli et al. 2009). To consider random variables, statistical sensitivity has been developed by several studies (Lee et al. 2011; Cho et al. 2014; Cho et al. 2016). After the relevance or influence measure is evaluated, essential variables can be determined according to the measure.

Especially, the number of samples used in variable selection—the efficiency of variable selection—is an important factor that should be carefully taken care in high-dimensional problems. Researchers (Székely et al. 2007, Cook 2000, Zhao et al. 2013) developed new influence diagnostic scores for efficient variable selection using a small number of samples. To ensure more stable result even with a few samples, there have been researches using the model-based method. Welch et al. utilized the likelihood of GPR as the criteria for variable screening (Welch et al. 1992). In the method, essential variables are detected by adding candidate variables one by one to the GPR model, and the variable which causes the highest improvement of the likelihood is selected. It was successfully applied to a 20-dimensional problem with less than 50 samples. Gaussian process (GP) classification (Rasmussen, 2006) calculates posterior using prior and given pre-labeled samples and makes small posterior smaller and large posterior larger. Hence, GP classification classifies relevant and irrelevant variables more distinctively. Recently, Wu et al. developed partial metamodel-based optimization utilizing radial basis function model with sensitive variables, which finally aims to obtain the optimal point with a reduced number of function evaluations (Wu et al. 2018). The method enhances the efficiency of the interwoven process of metamodeling and optimization by decomposing the space and focuses on the search area near the optimal point.

Throughout the literature survey, several aspects are found to be improved. First, several variable selection methods require well-positioned samples using delicate experimental design (Jin et al. 2003; Joseph et al. 2015; Joseph and Hung 2008; Pronzato and Walter 1988; Stein 1987; Beck and Guillas 2016; Ko et al. 1995; Gorodetsky and Marzouk 2016), or are coupled with specific experimental design methods (Hayter 2012; Moon et al. 2012). However, we may encounter samples that are not well-located. Second, the relevance measure that cannot identify the effect of interaction terms of variables on output response (Qi and Zhang 2001) could cause faulty variable selection in highly nonlinear problems. To capture the interaction term using Sobol' indices, enormously large number of samples (e.g., millions) could be required (Homma and Saltelli 1996; Sobol 2001). Statistical sensitivity method has the same drawback of requiring too many samples (Cho et al. 2014; Cho et al. 2016; Lee et al. 2011). Third, variable selection through influence diagnostic scores could produce a different selection of variables due to the fluctuation of influence diagnostics scores depending on the number of samples or locations of samples (Székely et al. 2007; Cook 2000; Zhao et al. 2013). Fourth, GP classification (Rasmussen, 2006) requires pre-labeled samples—the

predetermined data to which group they belong in classification (Guyon and Elisseeff 2003; Li et al. 2017; Chandrashekar and Sahin 2014). Finally, some modeling methods sacrifice the model accuracy, and the sensitivity analysis is performed with the intermediate metamodel to boost the optimization efficiency (Wu et al. 2018).

Therefore, this paper proposes an integrated variable selection and surrogate modeling algorithm that can cope with aforementioned difficulties—high dimensionality and nonlinearity, insufficient samples, arbitrary sample quality, the existence of interaction terms, and unlabeled data. The proposed method carries out the variable selection and surrogate modeling simultaneously without sensitivity analysis using as less number of samples as possible. In the proposed method, GPR is adopted for the surrogate modeling method, and variable subsets are evaluated with conservative multi-criteria. Variables that result in distinctly superior model performance are selected as the essential model inputs. Sequential adaptive sampling is carried out to avoid sample overuse. A conservative stopping criterion is developed to prevent premature stop of variable selection loop. For variable selection, two kinds of machine learning techniques are utilized for the proposed algorithm, clustering (Jain et al. 1999; Bouguettaya et al. 2015) and the Wrapper method (Kohavi and John 1997) since there is a functional similarity between data-driven modeling and physics-based modeling framework (Sun and Sun 2015; Solomatine and Ostfeld 2008; Bessa et al. 2017). Clustering is adopted to avoid ambiguous selection criteria because it can bisect data into distinctly different groups without a label. Especially agglomerative hierarchical clustering is adopted for the problem characteristic (Bouguettaya et al. 2015). The Wrapper method is also utilized due to its fitness for a variable selection in high uncertainty circumstances caused by a small number of samples. The Wrapper method determines the influence of a certain variable by the model performance formulated with the variable.

The organization of remaining parts of this paper is as follows. Section 2 reviews previous related researches. In Section 3, the proposed method is described in detail with an illustrative example. Section 4 presents two test examples to verify the performance of the proposed method. All results are summarized in Section 5.

2 Technological background

In this section, Sobol' indices and GPR model are briefly revisited. Sobol' indices and GPR are utilized as a benchmark to check variable selection accuracy and as surrogate modeling method, respectively, in this study. Specifically, the GPR model performances—marginal loglikelihood and integrated posterior variance—are core indices for variable selection criterion in the proposed method, and integrated posterior variance is also used for the experimental design criteria.

2.1 Sobol' indices

Sobol' indices is a global sensitivity index based on variance decomposition with Monte Carlo simulation. If the number of samples is enough, Sobol' indices can identify the accurate influence of input on output response for any type of functions such as highly nonlinear and high-dimensional problems with interaction terms. The basic concept of Sobol' indices is to decompose total variance to each variance caused by individual input and combinations of inputs (Homma and Saltelli 1996; Sobol 2001). When $\mathbf{I}$ is the unit interval [0, 1], $\mathbf{I}^n$ is the n-dimensional unit hypercube, and let us consider a function $f(\mathbf{x})$ with $\mathbf{x} = [x_1, x_2, \cdots, x_n] \in \mathbf{I}^n$ which can be formulated as

$$f(\mathbf{x}) = f(x_1, x_2, \cdots, x_n) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i<j}^{n} f_{ij}(x_i, x_j) + \cdots + f_{12\cdots n}(x_1, x_2, \cdots, x_n) \quad (1)$$

where n is the input dimension. Equation (1) is called ANOVA representation of $f(\mathbf{x})$ if

$$\int_0^1 f_{ij\ldots h}(x_i, x_j, \cdots, x_h)\, dx_i dx_j \cdots dx_h = 0 \text{ for } 1 \le i < j < \cdots < h \le n. \quad (2)$$

Equation (2) satisfies

$$f_0 = \int f(\mathbf{x}) d\mathbf{x}, \quad (3a)$$

$$f_i(x_i) = \int f(\mathbf{x}) \prod_{a \ne i} dx_a - f_0, \quad (3b)$$

$$f_{ij}(x_i, x_j) = \int f(\mathbf{x}) \prod_{a \ne i,j} dx_a - f_0 - f_i(x_i) - f_j(x_j), \quad (3c)$$

and so on. The total variance of f is defined as

$$D = \int f^2(\mathbf{x}) d\mathbf{x} - {f_0}^2 \quad (4)$$

which can be calculated as the sum of partial variances as

$$D = \sum_{i}^{n} D_i + \sum_{i<j}^{n} D_{ij} + \cdots + D_{12\cdots n} \quad (5)$$

where the partial variance is calculated as

$$D_{ij\cdots h} = \int f_{ij\cdots h}^2 dx_i dx_j \cdots dx_h \text{ for } 1 \le i < j < \cdots < h \le n \quad (6)$$

Using (5) and (6), Sobol' indices is defined as

$$S_{ij\cdots h} = \frac{D_{ij\cdots h}}{D} \text{ for } 1 \le i < j < \cdots < h \le n \quad (7)$$

2.2 Gaussian process regression

GPR is one of the most commonly used methods for data-driven modeling (Sun and Sun 2015). Although neural network (NN) is more widely used in data-driven modeling framework, GPR is better suited for computationally expensive cases that cannot provide a large number of samples. There are more reasons why GPR is chosen as a modeling method in this study. The first reason is that GPR is a well-formulated regression method that can cope with noise that results from the effect of removed variables on the original function value. Since the proposed algorithm uses a selective variable subset, the modeling method should be able to handle the noise. The second reason is that the marginal loglikelihood and posterior variance of GPR can be utilized as the variable selection measure.

GPR formulas are constructed as following procedure (Quiñonero-Candela and Rasmussen 2005; Rasmussen 2006; Bastos and O'Hagan 2009; Oakley and O'Hagan 2004). Random function $f(\mathbf{x})$ with zero mean and covariance function k with a hyperparameter set θ in a specified point $\mathbf{x}$ is expressed as

$$f(\mathbf{x}) \sim GP\left(0, k\left(\mathbf{x}, \mathbf{x}'; \theta\right)\right). \tag{8}$$

Assume that there are given m training data, that is, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m]^T \in \mathbb{R}^{m \times n}$ is a set of training inputs of m observations where each $\mathbf{x}_i \in \mathbb{R}^n$ is the ith training input. Training output $\mathbf{y}$ is a noisy version of $f(\mathbf{X})$ defined by $\mathbf{y} = f(\mathbf{X}) + \varepsilon$ where $f(\mathbf{X})$ is the latent function values of Gaussian process since the true noisy-free function value cannot be known, and $\varepsilon(i)$ is an independent and identically distributed Gaussian noise. When the signal noise variance is σ^2, the joint distribution of $\mathbf{y}$ and posterior prediction $\mathbf{f}_*$ on new input $\mathbf{X}_*$ follows multivariate normal distribution as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right) \tag{9}$$

where $\mathbf{I}$ is the identity matrix and K is defined with $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{v}$ and $\mathbf{XX} = \{\mathbf{xx}_j\}_{j=1}^{w}$ as

$$K(\mathbf{X}, \mathbf{XX}) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{xx}_1) & \cdots & k(\mathbf{x}_1, \mathbf{xx}_w) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_v, \mathbf{xx}_1) & \cdots & k(\mathbf{x}_v, \mathbf{xx}_w) \end{bmatrix}. \tag{10}$$

Based on Bayesian approach, the posterior prediction output $\mathbf{f}_*$ on the new input points $\mathbf{X}_*$ can be obtained with conditioning given training data as

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim N\left(\overline{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)\right) \tag{11}$$

with the posterior mean as

$$\overline{\mathbf{f}}_* = E(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{y}) = K(\mathbf{X}_*, \mathbf{X})\left[K(\mathbf{X}_*, \mathbf{X}) + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{y} \tag{12a}$$

and posterior covariance as

$$\mathrm{cov}(\mathbf{f}_*) = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})\left[K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}\right]^{-1} K(\mathbf{X}, \mathbf{X}_*). \tag{12b}$$

Next, a function $g(\mathbf{x})$ of nonzero mean which incorporates $f(\mathbf{x})$ with polynomial function $h(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}$ is considered as

$$g(\mathbf{x}) = h(\mathbf{x})^{\mathrm{T}} \boldsymbol{\beta} + f(\mathbf{x}) \tag{13}$$

where $h(\mathbf{x}) \in \mathrm{R}^{p \times 1}$ is a basis function consisting of p polynomial basis such as 1, x_1, x_2,..., x_n, x_1^2, x_2^2,..., x_n^2. We obtain the posterior output on the new input $\mathbf{X}_*$ conditioning training data as

$$\mathbf{g}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim N\left(\overline{\mathbf{g}}_*, \mathrm{cov}(\mathbf{g}_*)\right) \tag{14}$$

where the posterior mean is

$$\overline{\mathbf{g}}_* = \mathbf{H}_*^{\mathrm{T}} \hat{\boldsymbol{\beta}} + \mathbf{K}_*^{T}\left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1}\left(\mathbf{y} - \mathbf{H}^{\mathrm{T}} \hat{\boldsymbol{\beta}}\right), \tag{15a}$$

and posterior covariance is

$$cov(\mathbf{g}_*) = \mathrm{cov}(\mathbf{f}_*) + \left(\mathbf{H}_* - \mathbf{H}\left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{K}_*\right)^{\mathrm{T}} \left(\mathbf{H}\left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{H}^{T}\right)^{-1} \left(\mathbf{H}_* - \mathbf{H}\left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{K}_*\right) \tag{15b}$$

where $\mathbf{H} = [h(\mathbf{x}_1), h(\mathbf{x}_2), \cdots]^T$, $\mathbf{H}_* = [h(\mathbf{x}_{1*}), h(\mathbf{x}_{2*}), \cdots]^T$, $\mathbf{K}_* = K(\mathbf{X}_*, \mathbf{X})$, and $\mathbf{K} = K(\mathbf{X}, \mathbf{X})$ with $\mathrm{cov}(\mathbf{f}_*)$ in (12b).

For the parameter estimation, $\boldsymbol{\beta}$ can be estimated with σ^2 and θ as

$$\hat{\boldsymbol{\beta}}(\theta, \sigma^2) = \left(\mathbf{H}^{\mathrm{T}}\left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^{\mathrm{T}}\left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{y} \tag{16}$$

where σ^2 is the noise variance and θ is a hyperparameter vector of covariance function. The marginal likelihood is the integral of the likelihood times prior over the function values as

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f}. \tag{17a}$$

Since solving (17a) is intractable, plausible analytic solutions to solve (17a) have been suggested (Quiñonero-Candela and Rasmussen 2005; Rasmussen 2006; Bastos and O'Hagan 2009). In this study, we propose to use optimization to obtain the marginal likelihood with hyperparameters as

$$\left\{\hat{\theta}, \hat{\sigma}^2\right\} = \arg\max_{\theta,\sigma^2} p(\mathbf{y}|\mathbf{X}) \tag{17b}$$

where the loglikelihood with m observations can be approximated as (Rasmussen 2006)

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\left(\mathbf{y}-\mathbf{H}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\mathbf{K}+\sigma^2\mathbf{I}\right)^{-1} \left(\mathbf{y}-\mathbf{H}\hat{\boldsymbol{\beta}}\right) - \frac{m}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}+\sigma^2\mathbf{I}|. \tag{17c}$$

Hence, (17b) can be solved using (16) and (17c).

Predictive posterior variance $c(\mathbf{x}|\mathbf{X})$ is the same as $cov(\mathbf{g}_*)$ in (15b) except for replacing $\mathbf{X}_*$ with specified single design input $\mathbf{x}_*$. Then, the integrated posterior variance can be calculated as

$$IVAR = \int_{\mathbf{x}_*\in\chi} c(\mathbf{x}_*|\mathbf{X})d\mu \tag{18}$$

where μ is the PDF of $\mathbf{x}$ in space χ. This indicates μ-weighted integration over the design space χ. Through this process, two model performance indices for variable selection, marginal loglikelihood in (17c) and integrated posterior variance in (18), are obtained.

3 Proposed method

The main concept of the proposed method is that the marginal likelihood and integrated posterior variance are used as variable selection criteria. As mentioned in Introduction, the marginal likelihood was used as a variable screening criterion by Welch et al. (Welch et al. 1992), and the integrated posterior variance was reported as the model uncertainty measure by Gorodetsky et al. (Gorodetsky and Marzouk 2016). Hence, the marginal likelihood and the integrated posterior variance can be good measures to check the importance of input variables. Moreover, the integrated posterior variance is utilized for the sequential experimental design criteria at the same time. The following sections explain the proposed algorithm. Section 3.1 shows the basic concept of the proposed method and Sections from 3.2 to 3.5 explain the proposed algorithm in detail. Section 3.6 summarizes the proposed algorithm with a flowchart and Section 3.7 illustrates the proposed process with a simple mathematical example.

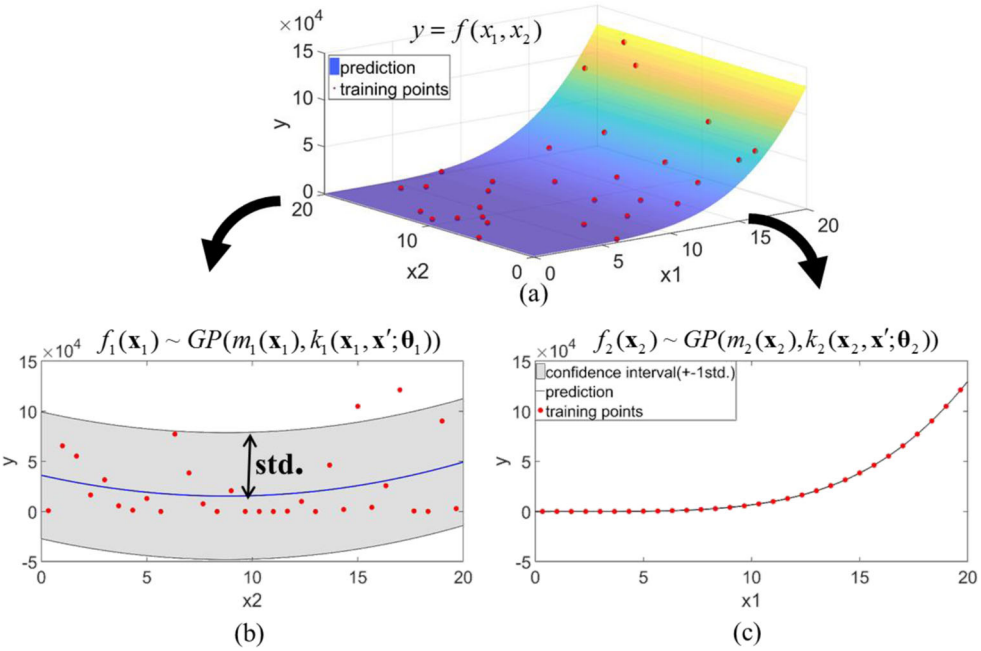

Fig. 1 **a**–**c** The concept of GPR and essential variable selection

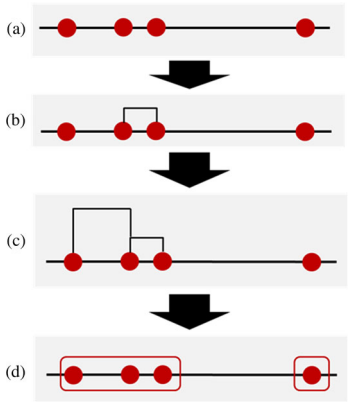

Fig. 2 **a**–**d** The concept of agglomerative hierarchical clustering

3.1 Basic concept of the proposed algorithm

The proposed algorithm distinguishes between essential and unnecessary variables using the Wrapper method (Kohavi and John 1997). The Wrapper method actually creates a surrogate model with a selected variable subset and determines the importance of variables based on the model performance of the surrogate model. When the performance of a GPR model with a certain variable set is clearly superior to models with other variable sets, variables in the set are significantly relevant to the output response and selected as essential variables. The model performance is quantified with (17c) and (18) after the GPR model is built.

Figure 1 visualizes the main concept of the proposed algorithm. Since the functional change almost depends on x_1 and x_2 has little effect as shown in Fig. 1a, x_2 is dispensable for the function. Figure 1b, c shows the function f when it is projected on x_2-y and x_1-y plane, respectively. The deviation of the GPR model with x_2 in Fig. 1b is much larger than the one in Fig. 1c. Hence, it can be seen that f is a function of mainly x_1, not x_2.

The intuitive concept of the proposed algorithm can be more comprehensively explained using (17c) given in the form of normal distribution. The first term of (17c) explains the L2-norm which is a discrepancy between true observation $\mathbf{y}$ and $\mathbf{H}\boldsymbol{\beta}$. The last term of (17c) is the log-determinant of the GPR model that means how much data is scattered from the regression. Because $\mathbf{H}\boldsymbol{\beta}$ is formulated with low-order polynomials, it is not enough to explain highly nonlinear and noisy response. Therefore, the covariance function handles the remaining elaborate scatter of data. If the data is largely scattered from the prediction, the marginal likelihood becomes small, and the variance becomes large. This is why the marginal likelihood and the integrated variance can be indicators that show whether the data can be well described with essential variables. These two indicators with various variable subsets can be divided distinctly into superior and inferior groups. The validity of two GPR model performances for the variable selection will be demonstrated in Section 3.7 in more detail using an illustrative mathematical example.

3.2 Variable selection algorithm

The subset selection process of the proposed method consists of three core methods: (1) the Wrapper method, (2) agglomerative hierarchical clustering, and (3) forward greedy search. As mentioned in Section 3.1, the Wrapper method is used to measure

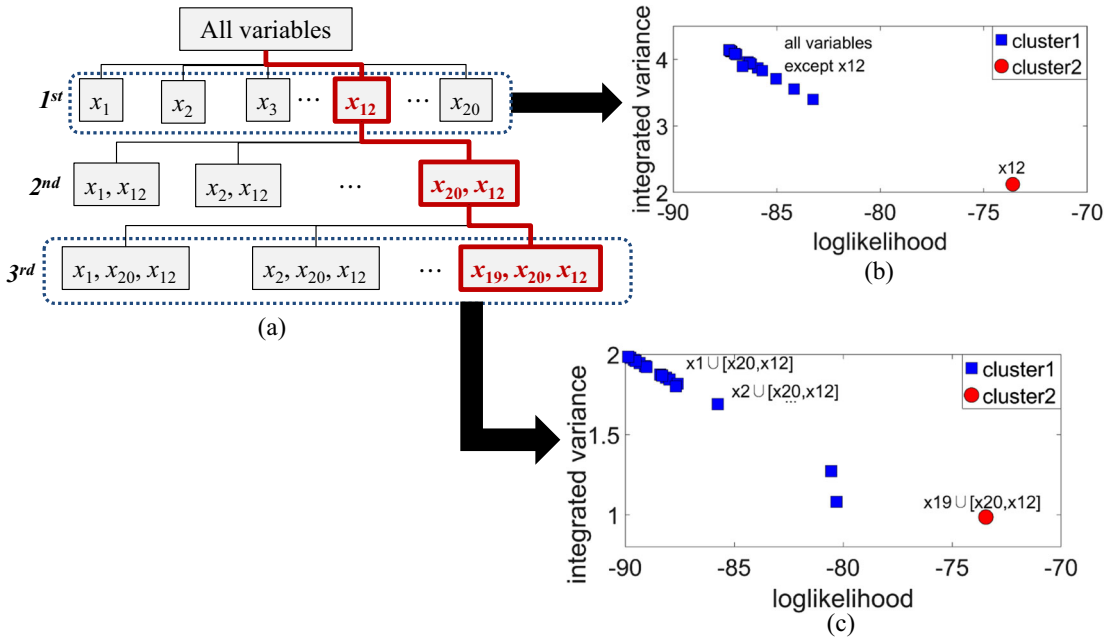

Fig. 3 **a**–**c** The concept of the proposed variable selection algorithm with a 20-dimensional example

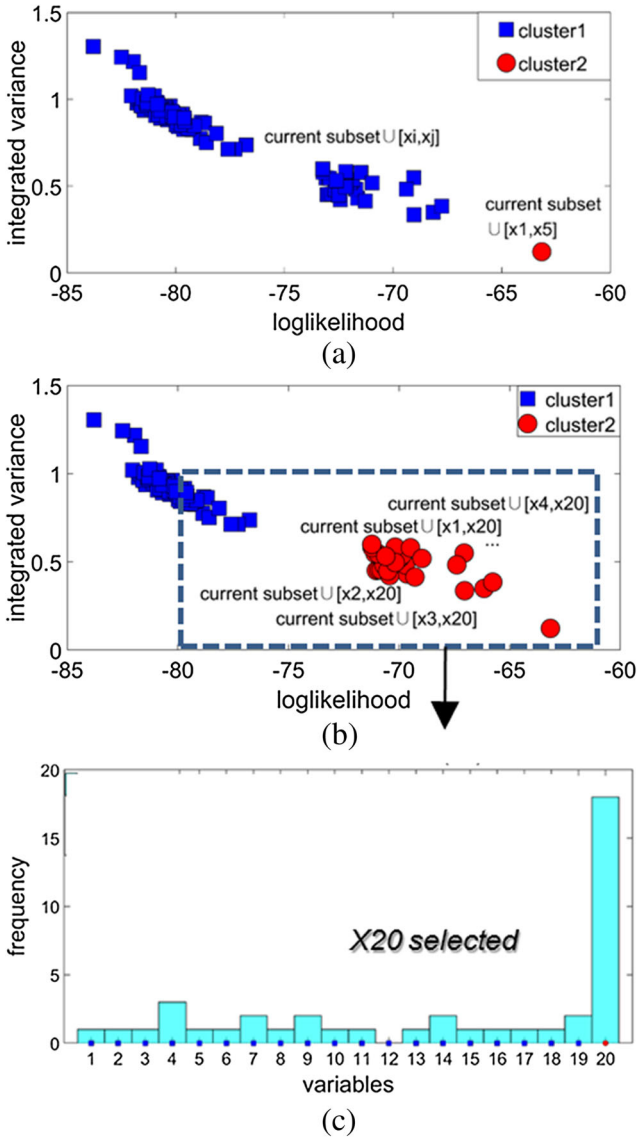

Fig. 4 **a**–**c** Two-variable addition method for variable selection

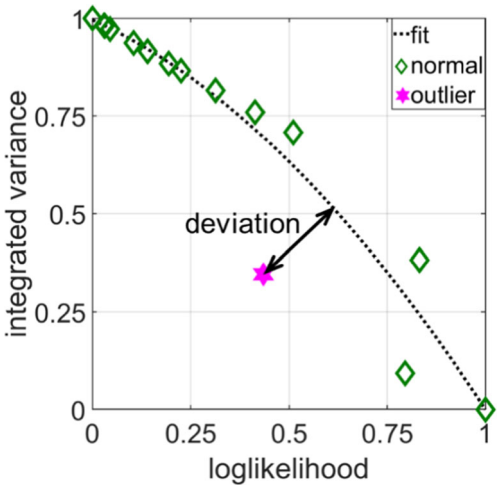

Fig. 5 Outlier detection in the normalized space of GPR model performances

the importance of variable sets. For the clustering method, agglomerative hierarchical clustering (Bouguettaya et al. 2015) is used to detect essential variables as shown in Fig. 2.

In this study, the number of the final clusters is set to two for data bisection, and Euclidean distance and single linkage are adopted as the distance metric and for linkage criterion, respectively. Single linkage criterion is the metric between two groups of **A** and **B** expressed as

$$\text{Linkage} = \min(\text{dist}(\mathbf{a}, \mathbf{b}) : \mathbf{a} \in \mathbf{A}, \mathbf{b} \in \mathbf{B}) \tag{19}$$

where dist(**a**,**b**) is Euclidian distance between **a** and **b**. Agglomerative hierarchical clustering links the closest data or clusters sequentially based on the linkage metric in (19) until the number of the clusters becomes 2.

Forward greedy search is a subset selection algorithm based on exhaustive search, which means that it finds the best subset by selecting variables one by one sequentially. Figure 3 shows the concept of the proposed variable selection algorithm combining three methods using a 20-dimensional example. As shown in Fig. 3a, forward greedy search is utilized to generate candidate variable sets in each iteration. For each variable, the Wrapper method is used to score the subset importance based on GPR model performances as shown in Fig. 3b, and agglomerative hierarchical clustering is used to distinguish the outstanding variable set which is x_{12} in Fig. 3b. In the second iteration, the forward greedy search is again applied to test all the combinations of two-variable sets with the selected x_{12}. In this way, the variable selection process continues until the GPR model obtained with selected variables only satisfies predefined accuracy.

Different from general greedy forward search that identifies the best subset using optimization, the proposed algorithm utilizes clustering as shown in Fig. 3b, c. Clustering can bisect data into definitely two different clusters. If there exists only one subset in the best cluster, it is classified as an important variable set. In addition, two model performance measures are utilized in the proposed method so that the important variable set which has the maximum of marginal loglikelihood and minimum of integrated posterior variance at the same time is selected. This conservativeness reduces the possibility of faulty variable selection.

Last but not least, the proposed algorithm utilizes two-variable addition when a definitely outstanding variable set does not exist due to strong interaction terms as shown in

Fig. 4. This two-variable addition is necessary for two reasons: (1) there are some variables that have high sensitivities when combined with other variables such as interaction terms, and (2) there can be two variables with similar sensitivities, and one variable cannot be distinguished from the other as an outstanding one. After performing two-variable addition, if there exists an outstanding set such as $x_{12} \cup [x_1, x_5]$ as shown in Fig. 4a, the variables in the set are classified into the important variable set. However, if there is no outstanding combination of two variables as shown in Fig. 4b even after two-variable addition, appearances of all the variables that belong to the best cluster (the box of Fig. 4b) are counted. If a variable repeatedly appears in the subsets of the best cluster as shown in Fig. 4c, it is classified as an essential one (x_{20} in Fig. 4b, c). If there is still no outstanding variable set even after clustering using two-variable addition, the algorithm performs sequential sampling to increase the accuracy of GPR model. Even if two-variable addition method can be easily extended to three or n-variable addition method, the number of variables to be added is restricted to 2 since the computational cost exponentially increases as the number increases.

Since the proposed method starts with a relatively small number of initial samples, it is possible that the wrong variables could be selected. In other words, a lack of samples or sample quality makes GPR model performance indices unstable which leads to fluctuation of variable importance ranking. Therefore, to avoid possible erratic results, a definitely outstanding variable must be selected that will be performed after outlier detection of the model performance in the proposed method.

3.3 Outlier detection

High deviation in the scatter plot of marginal loglikelihood and integrated variance means that the input sample quality or amount is not enough to select an outstanding variable. As shown in Fig. 5, if there exist outliers, one model performance index is improved while the other becomes worse at the outlier points. These outliers can be estimated using deviation from the regression line to the candidate outlier point, and if the deviation is larger than a target number, the point is identified as an outlier. In this study, the target number is set to 0.15 from the 2nd order regression in the normalized space as shown in Fig. 5. Outliers are detected when a GPR model is not accurate enough to be used for variable selection. Hence, once outliers are encountered, a sequential sampling that will be explained in the next section needs to be performed to increase the accuracy of the model.

3.4 Sequential sampling criterion for GPR model

In the proposed method, Latin hypercube sampling (LHS) that is a moderate level of sampling is utilized for initial sample generation in GPR modeling. However, there are two cases where the proposed algorithm performs sequential sampling in addition to the initial samples: (1) when an outlier exists in the model performances and (2) when accuracy of the model built with the current subset does not meet the target value, which means all of the important variables are not selected yet. For sequential sampling criterion, the method of a previous research (Gorodetsky and Marzouk 2016) that minimizes integrated posterior variance (IVAR) for GPR sampling is adopted in this study. In their research, it was explained that the IVAR criterion is equivalent to an expected integrated mean squared error (IMSE) of the posterior mean. The integrated posterior variance can be calculated using (18). According to the IVAR criterion, a new point $\mathbf{x}_a$ for sequential sampling can be obtained as

$$\mathbf{x}_a = \underset{\mathbf{x}\in\mathbf{U}}{\operatorname{argmin}} \int c(\mathbf{x}|\mathbf{X}) d\mu \approx \underset{\mathbf{x}\in\mathbf{U}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} c(\mathbf{x}|\mathbf{X}) \tag{20}$$

where $\mathbf{U}$ is the feasible design space and N is the number of Monte Carlo samples. If the integration in (20) is computationally expensive, $\mathbf{x}_b$ can be used to find a new sample point instead of (20) as

$$\mathbf{x}_b = \underset{\mathbf{x}\in\mathbf{U}}{\operatorname{argmax}}\, c(\mathbf{x}|\mathbf{X}) \tag{21}$$

which searches for a point of the highest posterior variance.

3.5 GPR model accuracy measure

To build a GPR model with reduced dimension as explained in Section 2.2, $\mathbf{X}$, an $n \times m$ matrix of current input samples where n denotes the number of sample and m denotes the dimension, is reduced to selected columns as $\mathbf{X} = \mathbf{X}_{(:,\,\mathbf{subset})}$ where subset denotes the selected variables. At the end of each iteration, the accuracy of the GPR model built with the subset is calculated using cross validation error (CVE) with the normalized root mean squared error ($n\text{–}rmse_{\mathrm{cve}}$) that is expressed as

$$n\text{–}rmse_{\mathrm{cve}} = \frac{1}{(y_{\max} - y_{\min})} \sqrt{\frac{1}{N_{\mathrm{samp}}} \sum_{i=1}^{N_{\mathrm{samp}}} {e_i}^2}, \qquad e_i = y_i - \hat{f}_{\sim i}(\mathbf{x}_i) \tag{22}$$

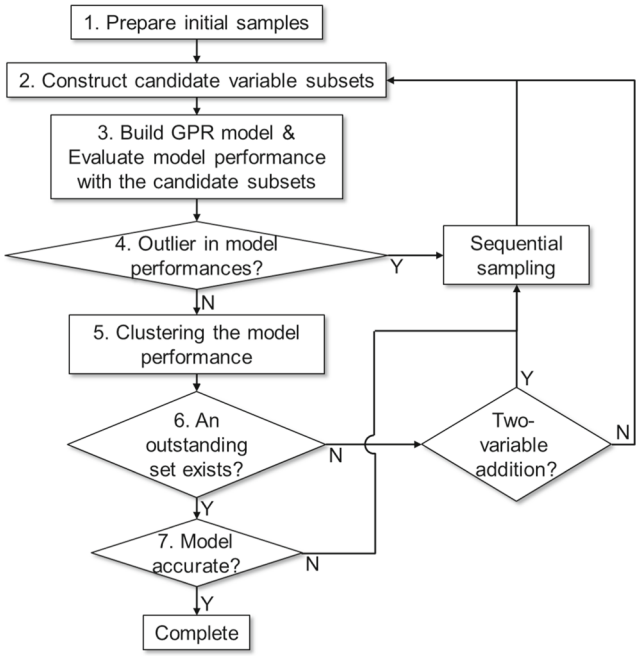

Fig. 6 Flowchart of the proposed algorithm

where y_{max} and y_{min} are the maximum and minimum output of the true training output, respectively, N_{samp} is the number of training samples, $\hat{f}_{\sim i}$ is the response from the surrogate model built without the i_{th} sample, and y_i is the true output evaluated at the i_{th} sample. As a stop criterion of the algorithm, $n\text{–}rmse_{cve}$ in (22) is utilized. However, to avoid premature convergence, two more samples are sequentially added to build additional GPR models once the GPR model satisfies the accuracy condition ($n\text{–}rmse_{cve} < 0.05$). Then, if all three GPR models satisfy the target accuracy, the algorithm stops.

When additional test samples are available, (22) is modified as

$$n\text{–}rmse = \frac{1}{(y_{max} - y_{min})}\sqrt{\frac{1}{N_{samp}}\sum_{i=1}^{N_{samp}} e_i^2}, e_i = y_i - \hat{f}(x_i) \quad (23)$$

where $\hat{f}$ is the response from the surrogate model built with all samples. Note that different from (22), y_{max} and y_{min} represent the maximum and minimum of the test output, respectively, and N_{samp} is the number of test samples in this case.

3.6 Overall algorithm

The overall process and pseudocode in Matlab form of the proposed algorithm explained in Sections 3.2~3.5 are presented in this section, and its flowchart is shown in Fig. 6.

Step 1: Generate $(2n + 1)$ initial samples using LHS where n is the problem dimension.
Step 2: Construct candidate variable subsets using forward greedy search. In the initial iteration, the candidate variable subsets are (x_1), (x_2),...(x_n). Continue to construct candidate variable subsets including selected variables.
Step 3: Using the constructed subsets in Step 2 and given samples, build GPR models and evaluate their model performances using the Wrapper method.
Step 4: Carry out outlier detection using the model performances obtained in Step 3. If outliers are detected, perform sequential sampling and return to Step 2.
Step 5: Perform clustering in the normalized space of two GPR model performances using the agglomerative hierarchical clustering.
Step 6: If there is no outstanding variable set identified after the clustering, perform two-variable addition in Step 2. If there is no outstanding variable set even after the two-variable addition, perform sequential sampling and return to Step 2.
Step 7: Evaluate the accuracy of the GPR model built with the selected variables and check the stopping criterion by Section 3.5.
Step 8: If the model satisfies the stopping criterion, then stop. Otherwise, repeat Steps 2 to 7 until the model accuracy is satisfied.

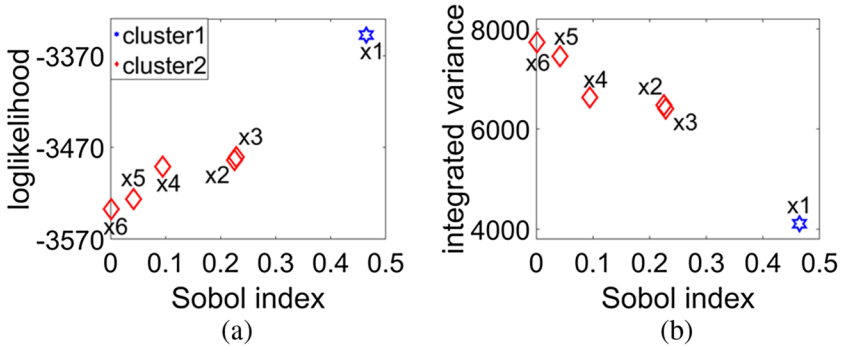

Fig. 7 **a** and **b** A monotonic relationship between two GPR model performances and Sobol' indices

Pseudocode of the proposed algorithm

```
1  Initialize the full variable set S, S←[1, 2, ... , n]
2  Initialize the selected variable subset F, F←Ø
3  Initialize the number of variables that will be added, N_add←1
4  Prepare initial samples, samp    % samp(i,j): i-th sample in j-th dimension
5  while(1) n-rmse_cve > crit
6        while(2) N_add < 3
7           A←all possible subsets of F^c with N_add elements % A(i,j):i-th subset in j-th dimension
8                 for(1) i =1:1:k    % k: number of rows of A
9                    F_cand(i,:)←F ∪ A(i,:)
10                   gpr←GPR(samp(:,F_cand(i,:)))    % GPR model construction with dim. F_cand(i,:)
11                   score(i , :) ←[loglikelihood(gpr), integ_variance(gpr) ] % model performances
12                end for(1)
13                if(1)  ∃ outlier, run sequential sampling; samp←[samp;new_samp]; N_add ←1
14                   continue % go to line7
15                end if(1)
16          run clustering
17                if(2)  ∃ an outstanding subset F_cand(mm,:), mm∈[1,2,…, k]
18                   F← F_cand(mm,:) ;break    % select the current best subset, go to line30
19                elseif there exist plural subsets in the best cluster & N_add =1
20                   N_add ← N_add +1; continue    % try two-variable addition, go to line7
21                elseif there exist plural subsets in the best cluster & N_add =2
22                      if(3)  ∃ a discriminately frequent variable var_best in the best
                        cluster
23                         F ← F ∪ var_best; break % go to line30
24                      else
25                         run sequential sampling; samp←[samp;new_samp]
26                         N_add ←1; continue % go to line7
27                      end if(3)
28                end if(2)
29       end while(2)
30    n-rmse_cve ← model accuracy with current subset F and current sample samp
31       if(4) n-rmse_cve > crit
32          run sequential sampling; samp←[ samp;new_samp]; N_add←1
33       end if(4)
34 end while(1)
```

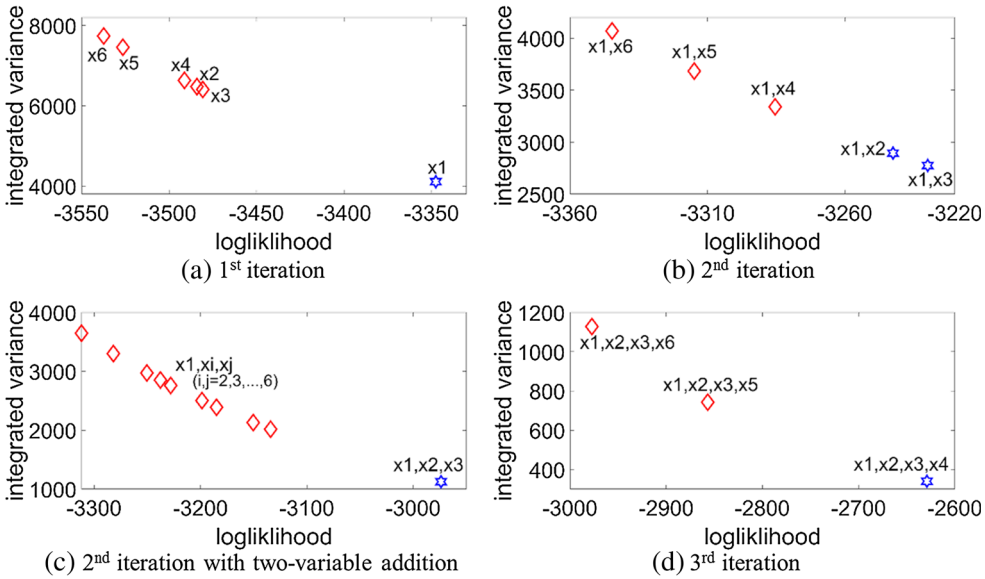

Fig. 8 **a**–**d** Variable selection process for illustrative example in (24)

3.7 An illustrative mathematical example for the proposed algorithm

For a better understanding of the proposed method and validity check of GPR model performances for variable selection, an illustrative example is utilized in this section given by

$$f(\mathbf{x}) = 200{x_1}^2 + 250x_2x_3 + 90{x_4}^2 + 60{x_5}^2 + 10x_6, x_i \in [0,1] \text{ for all } i = 1,\ldots,6. \quad (24)$$

This example includes an interaction term of x_2x_3 for which the two-variable addition explained in Section 3.5 is required. The number of initial samples is 600, which is 100 times the problem dimension, to remove the necessity of sequential sampling.

A previous research (Campolongo et al. 2007) developed a new sensitivity index and compared it with Sobol' indices. Similarly, we will show that there exists a monotonic relationship between GPR model performances and Sobol' indices using the illustrative example. As shown in Fig. 7a, b, there are monotonic relationships between the Sobol' indices and loglikelihood and Sobol' indices and integrated variance, respectively. This validates that two GPR model performances can replace Sobol' indices, which is very computationally expensive to obtain for variable selection.

Clustering results of successive iterations are shown in Fig. 8. It can be easily seen in Fig. 8a that x_1 is selected as an essential variable in the 1st iteration. In the 2nd iteration as shown in Fig. 8b, no one outstanding variable set exists due to the strong interaction between x_2 and x_3. Therefore, two-variable addition is performed, and x_2 and x_3 are included in the important variable set as shown in Fig. 8c. In the 3rd iteration, x_4 is added to the important variable set as shown in Fig. 8d. This variable selection process is summarized in Fig. 9. At the end of the 3rd iteration, the GPR model built with the subset (x_1,x_2,x_3,x_4) satisfies the target accuracy, that is, $n\text{-}rsme_{cve} < 0.05$, and thus the algorithm stops. Since the influence of x_5 and x_6 is ignorable as shown in Fig. 10, the GPR model with only four variables can replace the original 6-dimensional model without loss of accuracy.

This example intentionally eliminates the possibility of sequential sampling by using a sufficiently large number of initial samples which means there will be no outlier detected during the process. To test the proposed algorithm in the case of a small

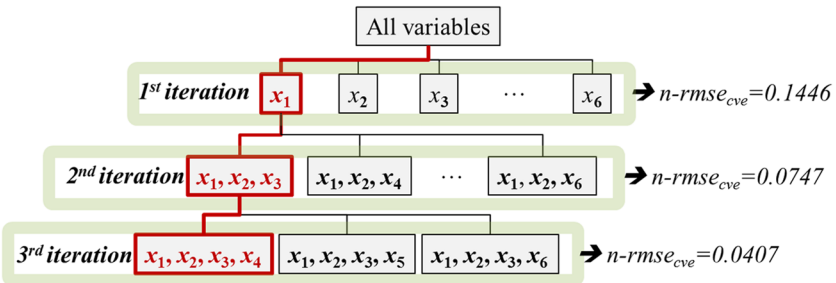

Fig. 9 Variable selection process and model accuracy in each iteration

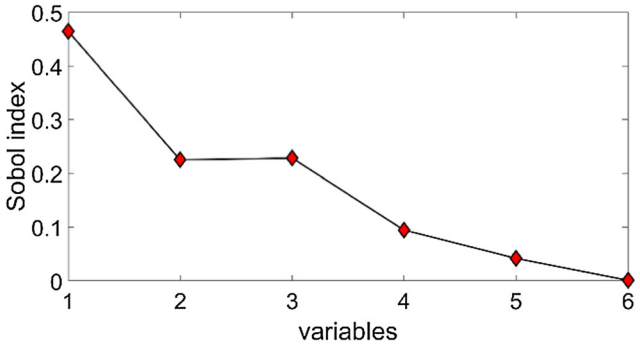

Fig. 10 Sobol' indices of illustrative example in (24)

number of initial samples, 13(= 2*n* + 1) initial samples by uniform random sampling and LHS are generated in 6-dimensional space as shown in Fig. 11a, b, respectively. From Fig. 11, it can be seen that LHS shows better space-filling capability compared with uniform random sampling meaning that sample quality by LHS is better than the one by uniform random sampling.

Figure 12a–c is the results generated from uniform random sampling and Fig. 12d is the result generated from LHS. Figure 12a shows there exists an outlier due to the poor sample quality whereas there is no outlier in Fig. 12c with the same sampling method since sufficient samples are used. It can be seen that sequential sampling reduces the deviation in Fig. 12a as shown in Fig. 12b and thus removes the outlier. This means that sequential sampling can alleviate instability of model performances caused by poor sample quality. This test also verifies the importance of initial sampling which is the reason why LHS is used in this study. As shown in Fig. 12d, there is no outlier in the GPR model generated using 13 samples with LHS.

4 Numerical examples

The usefulness of the proposed algorithm can be quantified by the accuracy of variable selection, final model accuracy, and the number of final samples used. The proposed algorithm is tested with two examples that have different properties. These

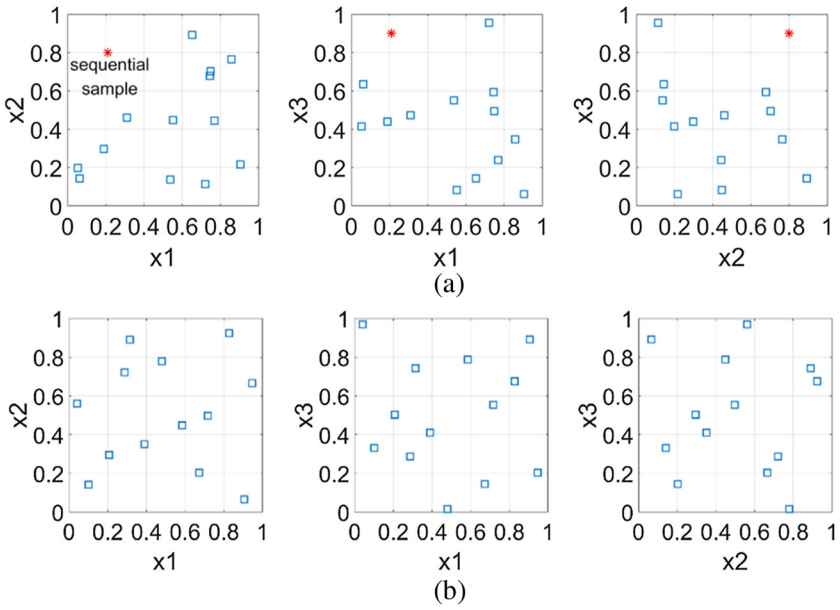

Fig. 11 Experimental design of **a** uniform random sampling and **b** LHS

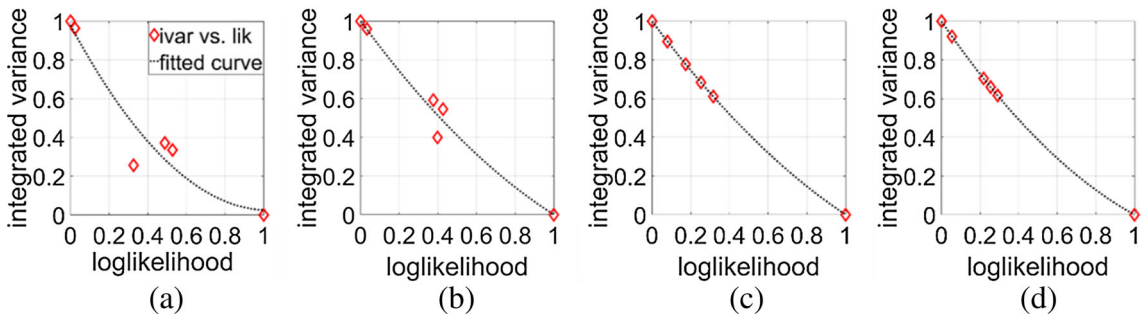

Fig. 12 Normalized GPR performances generated by **a** uniform random 13 samples, **b** uniform random 13 samples + 1 sequential sample, **c** uniform random 600 samples, and **d** Latin hypercube 13 samples

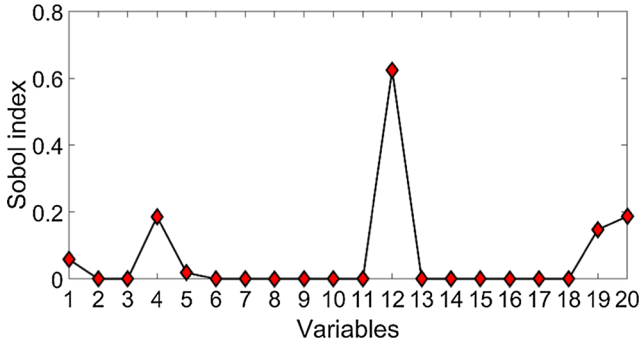

Fig. 13 Sobol' indices of the example in (25)

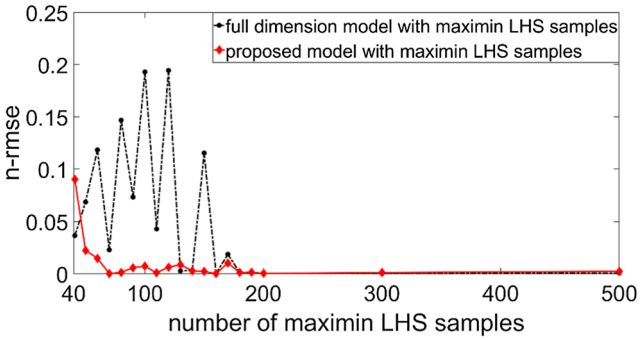

Fig. 15 Comparison of model accuracy between conventional and proposed methods of (25)

examples verify the excellent performance of the proposed algorithm with high model accuracy and with a small number of samples.

4.1 Mathematical example

Welch et al. performed variable screening for a 20-dimensional and highly nonlinear mathematical example given by (Welch et al. 1992)

$$f(\mathbf{x}) = \frac{5x_{20}}{1+x_1} + 5(x_4 + x_{20})^2 + x_5 + 40{x_{19}}^3 - 5x_{19} + 0.05x_2 + 0.08x_3 - 0.03x_6 + 0.03x_7 - 0.09x_9 - 0.01x_{10} - 0.07x_{11} + 0.25{x_{13}}^2 - 0.04x_{14} + 0.06x_{15} - 0.01x_{17} - 0.03x_{18}, x_i \in [-0.5, 0.5] \text{ for all } i = 1, \ldots, 20. \quad (25)$$

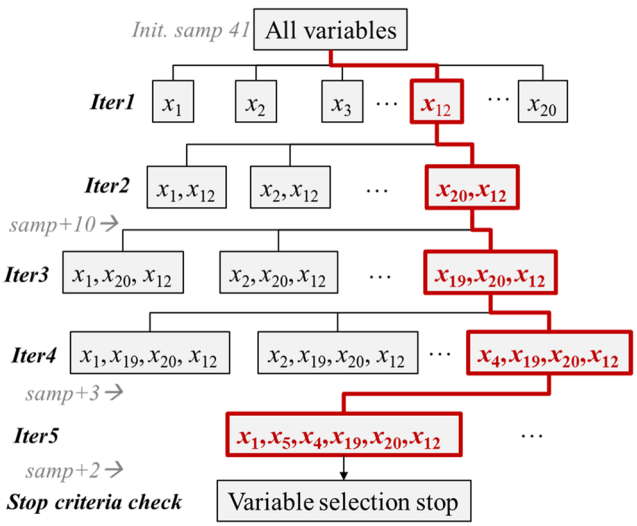

Fig. 14 Sequential variable selection process of (25)

from which it can be seen that there exist interaction terms as well as nonlinear terms. The Sobol' indices of variables of this example obtained using one million samples are shown in Fig. 13. As shown in Fig. 13, x_{12} is the most important variable, and x_{12}, x_4, x_{20}, x_{19}, x_1, and x_5 could be selected as essential variables.

The result of the proposed algorithm is summarized in Fig. 14 and Table 1. From the figure, it is shown that x_{12}, x_{20}, x_{19}, x_4, x_5, and x_1 are sequentially selected which is identical with the result of the Sobol' indices. It can also be seen that the variables with low Sobol' indices such as (x_1, x_5) are identified later than the ones with high indices such as (x_4, x_{12}, x_{19}, x_{20}). $n\text{–}rmse_{\text{cve}}$ of the GPR model built with the selected variables is lower than 0.05 with 54 samples; however, 55th and 56th samples are used to avoid premature convergence of model accuracy. Hence, total samples used are 56, which is only 0.0058% of those of the Sobol' indices, while the variable selection results are identical verifying that the proposed method selects essential variables accurately and efficiently.

It should be noted that the GPR model generated using the proposed method is much more accurate than the surrogate model of full dimensions especially when the number of samples used is small as shown in Fig. 15. The full-dimension surrogate model requires more than 150 maximin LHS samples to satisfy the model accuracy criterion while the proposed method requires only 56 LHS samples. *n–rmse* in Fig. 15 is

Table 1 Variable selection results of (25)

Iteration	Selected variables	$n\text{–}rmse_{\text{cve}}$	# of samples
1	12	0.2107	41
2	20,12	0.2043	41
3	19,20,12	0.1001	51
4	4,19,20,12	0.0640	51
5	1,5,4,19,20,12 (stop criteria check)	*0.0486*	54
6	1,5,4,19,20,12 (stop criteria check)	*0.0466*	55
7	1,5,4,19,20,12 (stop criteria check)	*0.0441*	56

The italic font indicates the stop criteria check step

Table 2 Function definition

Mode		Function	Value
Safety	Full frontal impact	G_1	Chest G
		G_2	Crush displacement
	40% offset impact	G_3	Brake pedal
		G_4	Footrest
		G_5	Left toepan
		G_6	Center toepan
		G_7	Right toepan
		G_8	Left instrument panel
		G_9	Right instrument panel
NVH		G_{10}	Torsion mode
		G_{11}	Vertical bending mode

calculated with additionally generated 100 test samples. In conclusion, the more accurate surrogate model is generated in this example using the proposed algorithm with less number of samples by removing unnecessary variables.

Table 3 Variable selection results of the engineering example

Iteration	Selected variables	n–$rmse_{cve}$	# of samples
1	22	0.1415	89
2	24,22	0.1254	89
3	18,24,22	0.1167	90
4	21,18,24,22	0.1006	92
5	1,21,18,24,22	0.0888	92
6	10,1,21,18,24,22	0.0760	92
7	9,19,10,1,21,18,24,22	0.0771	92
8	11,9,19,10,1,21,18,24,22	0.0667	103
9	12,11,9,19,10,1,21,18,24,22	0.0559	103
10	3,12,11,9,19,10,1,21,18,24,22	0.0554	103
11	16,23,3,12,11,9,19,10,1,21,18,24,22 (stop criteria check)	*0.0443*	103
12	16,23,3,12,11,9,19,10,1,21,18,24,22 (stop criteria check)	*0.0445*	104
13	16,23,3,12,11,9,19,10,1,21,18,24,22 (stop criteria check)	*0.0460*	105

The italic font indicates the stop criteria check step

4.2 Engineering example

Cho et al. carried out statistical variable screening using 11 functions with 44 variables which are nine vehicle safety performances and two noise, vibration, and harshness (NVH) performances as listed in Table 2 (Cho et al. 2016). Six variables (x_1~x_5 and x_8) are common variables for all 11 functions, and two variables (x_6 and x_7) are for safety only, and the other 36 variables are for NVH. All the 44 input variables correspond to the steel plate thickness of each part of the vehicle NVH model. As shown in Fig. 16, many variables in G_{11} have very similar and low Sobol' indices meaning that the indices are not easily distinguishable. For the reason, G_{11} is selected among 11 functions to validate the proposed algorithm.

The result of the proposed method is summarized in Table 3. In 13 iterations, 13 variables among 44 are selected as essential variables. This result is identical with variable selection by the Sobol' indices as shown in Table 4. The total number of samples used is 105 which is very efficient considering that conventional surrogate models with full dimension would use at least 132 samples ($3n$) as the initial sample.

The accuracy of the GPR model built with the selected variables is compared with the full-dimensional model in Fig. 17. Since the variables are not highly correlated and the original model is very accurate with only 100 samples in this example, the accuracy of the full-dimensional model is similar to the one of the reduced GPR model. However, since users do not know how many samples are required for the model in

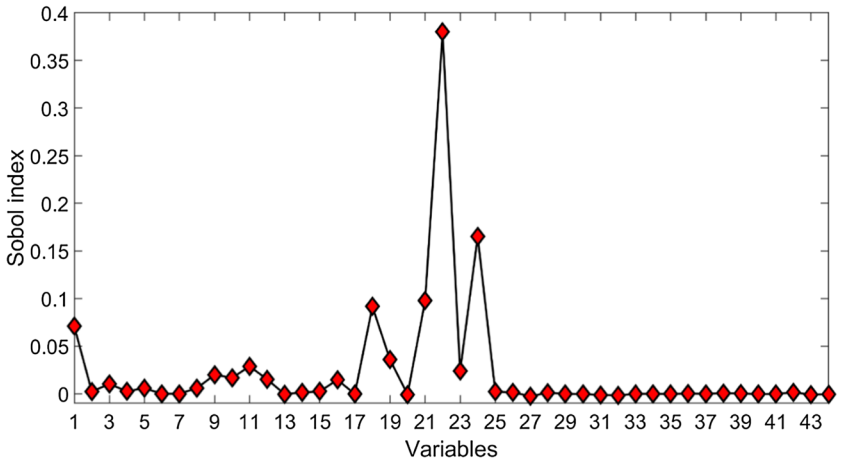

Fig. 16 Sobol' indices for G_{11} of engineering example

Table 4 Rank of Sobol' indices and variable selection results of engineering example

Sobol' indices	Variables
> 0.1	**22,24**
0.05~0.1	**18,21,1**
0.01~0.05	**3,9,10,11,12,16,19,23**
0.005~0.01	5,8
< 0.005	Remainder

The bold font indicates the selected important variables

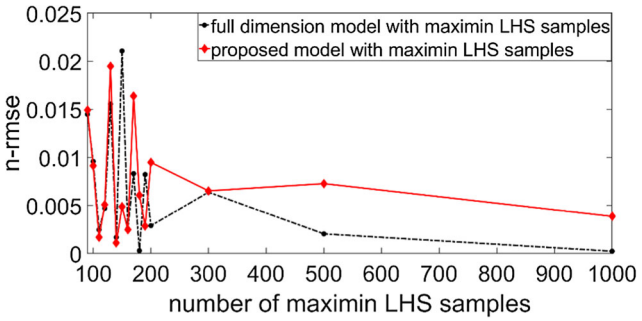

Fig. 17 Model accuracy comparison between the conventional and proposed model of the engineering example

advance, the proposed method can be regarded as a meaningful preprocess for general-purpose-use.

5 Conclusion

In this paper, we present an efficient variable selection methodology to overcome the curse of dimensionality in surrogate modeling. GPR is used for the surrogate modeling, and marginal loglikelihood and integrated posterior variance are used as measures to select essential variables according to the Wrapper method. Variables that induce high marginal loglikelihood and low integrated posterior variance in surrogate modeling are selected as essential variables. To find essential variables systematically, the greedy forward search has been modified and utilized. Agglomerative hierarchical clustering is adopted to distinguish essential variables and the others clearly. When essential variables cannot be identified clearly due to a limited number of samples in each algorithm loop, adaptive samples are sequentially added to increase the accuracy of GPR models. Major contribution of the proposed algorithm is that it effectively reduces required number of samples for surrogate modeling when the fraction of dispensable variables is high since the number of samples depends only on essential variables in the proposed method. Another advantage is that the algorithm is less dependent on the initial sample quality since sequential sampling is utilized. In addition, it can carry out both variable selection and surrogate modeling simultaneously. The proposed algorithm was verified with two numerical examples with high dimensions and challenging properties. A numerical study shows that with a much smaller number of samples variable selection accuracy is almost the same with the Sobol' indices, and the GPR model shows similar or better accuracy than the conventional full-dimension model.

Funding information This research was supported by the development of thermoelectric power generation system and business model utilizing non-use heat of industry funded by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea (No.20172010000830).

References

Bastos LS, O'Hagan A (2009) Diagnostics for Gaussian process emulators. Technometrics 51(4):425–438

Beck J, Guillas S (2016) Sequential design with mutual information for computer experiments (MICE): emulation of a tsunami model. SIAM/ASA Journal on Uncertainty Quantification 4(1):739–766

Bessa MA, Bostanabad R, Liu Z, Hu A, Apley DW, Brinson C, Liu WK (2017) A framework for data-driven analysis of materials under uncertainty: countering the curse of dimensionality. Comput Methods Appl Mech Eng 320:633–667

Bouguettaya A, Yu Q, Liu X, Zhou X, Song A (2015) Efficient agglomerative hierarchical clustering. Expert Syst Appl 42(5):2785–2797

Campolongo F, Cariboni J, Saltelli A (2007) An effective screening design for sensitivity analysis of large models. Environ Model Softw 22(10):1509–1518

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Cho H, Bae S, Choi KK, Lamb D, Yang RJ (2014) An efficient variable screening method for effective surrogate models for reliability-based design optimization. Struct Multidiscip Optim 50(5):717–738

Cho H, Choi KK, Gaul NJ, Lee I, Lamb D, Gorsich D (2016) Conservative reliability-based design optimization method with insufficient input data. Struct Multidiscip Optim 54(6):1609–1630

Cook RD (2000) Detection of influential observation in linear regression. Technometrics 42(1):65–68

Gorodetsky A, Marzouk Y (2016) Mercer kernels and integrated variance experimental design: connections between Gaussian process regression and polynomial approximation. SIAM/ASA Journal on Uncertainty Quantification 4(1):796–828

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hayter A (2012) Probability and statistics for engineers and scientists. Nelson Education

Helton JC, Johnson JD, Sallaberry CJ, Storlie CB (2006) Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab Eng Syst Saf 91(10–11):1175–1209

Homma T, Saltelli A (1996) Importance measures in global sensitivity analysis of nonlinear models. Reliab Eng Syst Saf 52(1):1–17

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Jin R, Chen W, Sudjianto A (2002) On sequential sampling for global metamodeling in engineering design. In ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers

Jin, R., Chen, W., and Sudjianto, A. (2003) An efficient algorithm for constructing optimal design of computer experiments. in ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers

Joseph VR, Hung Y (2008) Orthogonal-maximin Latin hypercube designs. Stat Sin 171–186

Joseph VR, Gul E, Ba S (2015) Maximum projection designs for computer experiments. Biometrika 102(2):371–380

Ko CW, Lee J, Queyranne M (1995) An exact algorithm for maximum entropy sampling. Oper Res 43(4):684–691

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324

Lee I, Choi KK, Noh Y, Zhao L, Gorsich D (2011) Sampling-based stochastic sensitivity analysis using score functions for RBDO problems with correlated random variables. J Mech Des 133(2):021003

Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2017) Feature selection: a data perspective. ACM Computing Surveys (CSUR) 50(6):94

Moon H, Dean AM, Santner TJ (2012) Two-stage sensitivity-based group screening in computer experiments. Technometrics 54(4):376–387

Oakley JE, O'Hagan A (2004) Probabilistic sensitivity analysis of complex models: a Bayesian approach. J R Stat Soc Ser B Stat Methodol 66(3):751–769

Pronzato L, Walter É (1988) Robust experiment design via maximin optimization. Math Biosci 89(2):161–176

Qi M, Zhang GP (2001) An investigation of model selection criteria for neural network time series forecasting. Eur J Oper Res 132(3):666–680

Quiñonero-Candela J, Rasmussen CE (2005) A unifying view of sparse approximate Gaussian process regression. J Mach Learn Res 6(Dec): 1939–1959

Rasmussen CE, Williams CK (2006) Gaussian process for machine learning, cambridge, MIT press

Saltelli A, Campolongo F, Cariboni J (2009) Screening important inputs in models with strong interaction properties. Reliab Eng Syst Saf 94(7):1149–1155

Shan S, Wang GG (2010) Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. Struct Multidiscip Optim 41(2):219–241

Sobol IM (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 55(1–3):271–280

Solomatine DP, Ostfeld A (2008) Data-driven modelling: some past experiences and new approaches. J Hydroinf 10(1):3–22

Stein M (1987) Large sample properties of simulations using Latin hypercube sampling. Technometrics 29(2):143–151

Sun NZ, Sun A (2015) Model calibration and parameter estimation: for environmental and water resource systems. Springer

Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794

Welch WJ, Buck RJ, Sacks J, Wynn HP, Mitchell TJ, Morris MD (1992) Screening, predicting, and computer experiments. Technometrics 34(1):15–25

Wu D, Hajikolaei KH, Wang GG (2018) Employing partial metamodels for optimization with scarce samples. Struct Multidiscip Optim 57(3):1329–1343

Zhao J, Leng C, Li L, Wang H (2013) High-dimensional influence measure. Ann Stat 41(5):2639–2667