Computational and Applied Mathematics (2024) 43:9
https://doi.org/10.1007/s40314-023-02519-z

A sequential designing-modeling technique when the input factors are not equally important

A. M. Elsawah[1,2,3] · Yi-An Wang[2] · Zhihan Chen[2] · Fatih Tank[4]

Received: 11 February 2023 / Revised: 8 September 2023 / Accepted: 3 November 2023 /
Published online: 7 December 2023

Abstract

The first thing springs to mind for understanding, forecasting, and improving the behavior of a complex system is a data-based model. This paper presents a sequential designing-modeling technique when the input factors do not have the same influence. The power of the combination of the design of experiments approach and modeling approach is investigated. The proposed technique adds the input factors to the process and designs and models them one after the other. At each step, one input factor is added based on its significance (impact), while each remaining input factor is set at its highest-influencing point (value). Ranking the factors in terms of significance and determining the point that has the highest effect for each factor are investigated. A comparison study between the new proposed sequential-stages technique (SeqST) and the classical single-stage technique (SinST) is given. The main results show that: (i) the performance of the SeqST is better than the performance of the SinST under different experimental conditions and scenarios, (ii) when there is a small number of training points in an experiment, there is a larger difference between the performance of the SeqST and the SinST than there is when there is a large number, (iii) when there are huge gaps between the importance of the factors in an experiment, there is a larger difference between the performance of the SeqST and the SinST than there is when there are small gaps, (iv) the SeqST has a much better performance using the correct order of the importance of the factors, and (v) the SeqST has a much better performance using a descending order of the numbers of the training points in the follow-up stages. In conclusion, for experiments with few trials and/or big gaps between the factors' importance, it is highly recommended to use

Communicated by Graçaliz Pereira Dimuro.

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.cn; a.elsawah@zu.edu.eg

[1] Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519087, China

[2] Department of Statistics and Data Science, Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519087, China

[3] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

[4] Department of Actuarial Sciences, Faculty of Applied Sciences, Ankara University, Cankaya, Ankara, Turkey

the SeqST with the ascending order of the factors' importance and a decreasing order of the numbers of training points in the follow-up stages. This study gives a benchmark that guide experimenters to effectively designing and modeling their experiments.

Keywords Design of experiment · Modeling · Sequential stages' experiment · Single-stage experiment · Latin hypercube design · Mass centers

Mathematics Subject Classification 62K05 · 62K15

1 Introduction

The first thing springs to mind for understanding, forecasting, and improving the behavior of complex experiments for real-life phenomena, industrial applications, and scientific investigations is a data-based model. Designing and modeling a studied experiment are the two key stages for this purpose. The significant purpose of the first stage, designing the experiment, is the selection of a representative dataset that provide precise information and correct understanding about the most significant features and behavior of the phenomenon under the experimentation (cf. Elsawah 2021a). Modeling the collected representative dataset, i.e., screening the relationship between input factors and their responses, is the second stage that can be used to estimate unknown parameters and predict the behavior of the studied phenomenon and thus guide the investigators to improve the inputs or experimental conditions for optimizing the corresponding outputs (cf. Elsawah 2021b). This logical idea is a classical methodology that is extremely used in computer and physical experiments (cf. Elsawah 2023a, b). For example, it is used in the industry in designing the process, reducing the process time, improving the quality of the products by reducing variability and increasing reliability, and reducing the overall costs (cf. Elsawah 2022a).

Efficient designing and modeling methods are able to capture maximum valuable (accurate) information about the behavior of a given experiment, and thus, an efficient model can be established based on the optimal representative dataset to screen the relationship between the inputs and their corresponding responses that can be used to estimate significant unknown parameters without bias and with minimum variance and forecast the future behavior of the studied phenomenon. Whereas non-efficient designing or/and modeling methods cannot produce useful and correct information nor provide accurate estimation or prediction (Elsawah 2022b). The practice demonstrated that effectively designing and modeling experiments are significant hard problems experimenters may face in many real-life applications. Despite the fact that many approaches have been offered, the challenge faced by the experimenters is still daunting.

The significant problem in improving the designing and modeling methods is that the researchers are improving the methods of each stage independently. On one hand, the idea of design of experiment approach (Fisher 1935) and the corresponding approaches and developments are used to improve the first stage, designing the experiments, and many efficient methods are given to optimally select representative datasets. On the second hand, the power of the modeling approach and its corresponding methods such as machine learning (Samuel 1959) are used to improve the second stage, modeling the experiments. However, these two approaches are complementary and not alternative and their power can be merged to support each other. The combination of design of experiment and modeling has recently attracted the attention of researchers (cf. Lujan-Moreno et al. 2018; Salmaso et al. 2022).

Even though there is obvious link between the design of experiment and modeling, there are surprisingly few papers on addressing the potential usefulness of a combination of the two concepts. For instance, Staelin (2003) used the principles of design of experiment to identify optimal or nearly optimal initial parameter settings in an example of support vector machines; Packianather et al. (2000) applied the Taguchi design approach to optimize the design parameters in an example of neural networks; Sukthomya and Tannock (2005), Ortiz-Rodriguez et al. (2006), and Balestrassi et al. (2009) all reached the conclusion that the design of experiment approach allows for gaining a profound understanding of the effects of parameters on the network performance and hence enables better parameter adjustments. The existing work compares or combines the two concepts in specific areas of interest or for specific problem investigations (cf. for example Mohamed et al. 2023; Prasath et al. 2021, 2022), but a paper producing a generalizable assessment of how the two methodologies can be applied jointly to develop a new efficient designing-modeling approach has not been put forward so far and the work in this topic is limited. Readers who are interested in learning more new approaches for designing or modeling experiments may refer to Sikirica et al. (2023), Iordanis et al. (2022), Zhang et al. (2022) and Elsawah (2017a, b).

Consider an explicit function for an experiment with p input factors $X_1, X_2, \ldots, X_p$ and only one output factor Y and the experimenter wants to estimate the true model $Y = F(X_1, X_2, \ldots, X_p)$ that gives the relationship between the p input factors and their corresponding responses. The classical modeling technique estimates the model $Y = F(X_1, X_2, \ldots, X_p)$ in one step based on a selected representative dataset that is an $n \times p$ data matrix by selecting n different values from the range of each input factor $X_i,\ i = 1, \ldots, p$. However, the accuracy of the approximate model $\widehat{Y} = \widehat{F}(X_1, X_2, \ldots, X_p)$ in many cases is not good, especially when there is no or little prior information about the true model. Therefore, the logical idea is that: *The weight of importance of each input factor needs to be taken into the consideration and a closer look at the sub-models between the most important input factors and their corresponding responses need to be investigated.* This paper presents a sequential designing-modeling technique (SeqST) that takes the weight of the importance of each input factor into consideration. The power of the combination of the sequential design of experiment approach and sequential modeling approach is investigated. The input factors are added to the proposed technique and modeled sequentially, one input factor is added at each stage, according to their importance (i.e., expected influence on the output), while each remaining input factor keeps fixed at a given point (value) that has the highest influence based on a prior knowledge or an initial experiment (cf. Sect. 3 for more details). Based on this simple introduction of the new proposed SeqST, the following logical questions may arise: *How to rank the importance of the input factors in order? How to find the point of the highest influence for each factor? What is the effect of the total number of training points on the performance of the SeqST? What is the effect of the number of training points in each stage on the performance of the SeqST? What is the effect of the order of the importance of the input factors on the performance of the SeqST? What is the effect of the gap between the importance of the input factors on the performance of the SeqST?* This paper tries to answer these interesting questions to investigate the performance of the proposed SeqST for different scenarios that give benchmarks to guide the experimenters to effectively designing and modeling their experiments. The power of the new proposed SeqST is measured by comparing its performance with the performance of the classical modeling technique, single-stage technique (SinST).

The rest of this paper is organized as follows. Section 2 gives the new proposed SeqST. Measuring the importance of each factor and finding the point with the highest influence for each factors are discussed in Sect. 3. Section 4 gives an illustrative example based on the discussions in Sects. 2 and 3. The performance of the new proposed SeqST is compared with the performance of the SinST using linear and non-linear models in Sect. 5. Section 6 gives further investigations for the performance of the proposed SeqST using different scenarios of the number of training points and the order of the importance of the input factors. We close through the conclusion and future work in Sect. 7.

2 The new proposed sequential stages designing-modeling technique

Consider an experiment with p input factors $X_1, X_2, \ldots, X_p$ and only one output factor Y and the experimenter wants to find the meta-model $\widehat{Y} = \widehat{F}(X_1, X_2, \ldots, X_p)$ that gives the relationship between the p input factors and their corresponding responses. This paper presents a step-by-step technique for incorporating design of experiment approach into modeling approach and adapting it to address some drawbacks of the existing techniques. Due to the limitation of the space and for a clear explanation, the new proposed SeqST uses the regression model from the modeling approach, which is the most basic strategy in the modeling approach and its success is more conducive to the proliferation of other advanced models. However, many different models can be used to extend this study. The new proposed SeqST is given by the following steps:

- ***Preparation stage:*** Rank the p inputs $X_1, X_2, \ldots, X_p$ according to their importance, i.e., influence on the output. Let $X_{1:p} \ggg X_{2:p} \ggg \cdots \ggg X_{p:p}$ is the corresponding importance order of the p input factors, where $X_{1:p}$ is the input with the highest importance and $X_{p:p}$ is the input with the lowest importance. Determine the most important level (value) of each input factor, i.e., the value for each factor that has the highest importance. Let $x^*_{1:p}, x^*_{2:p}, \ldots,$ and $x^*_{p:p}$ are the p highest influence levels of the p input factors $X_{1:p},\ X_{2:p}, \ldots,$ and $X_{p:p}$, respectively. It is worth mentioning that the importance (or influence) of the input factors and their levels that have the highest influences can be given based on expert knowledge or prior information by investigating an initial small experiment. If there is no prior information, Sect. 3 investigates a theoretical method to estimate the importance of each factor and the point with the highest influence for each factor.
- ***First designing-modeling stage:*** Generate the first-stage dataset (design) that is an $n_1 \times p$ data matrix $\mathbf{U}_1 = \left[\mathbf{D}_1, X^*_{2:p}, \ldots, X^*_{p:p}\right]$, where $\mathbf{D}_1 = \left(x^{(1)}_{1:p}, \ldots, x^{(n_1)}_{1:p}\right)^T$ is an optimal design from the experimental design viewpoint over the domain of the highest importance input factor $X_{1:p}$ and $X^*_{k:p} = (x^*_{k:p}, \ldots, x^*_{k:p})^T$ is a vector that all of its n_1 values are fixed to the highest importance level value $x^*_{k:p}$ of the kth input factor $X_{k:p}$ for $k = 2, \ldots, p$. Calculate the first-stage observed output vector via a physical experiment or exact output vector via a computer experiment, say $Y_1 = F(\mathbf{U}_1)$. Find the first-stage meta-model $\widehat{F_1}$ that is the approximate model for the relationship between the $\mathbf{D}_1$ in the first-stage design $\mathbf{U}_1$ and the corresponding first-stage observed output factor $Y_1 = F(\mathbf{U}_1)$.
- ***Second designing-modeling stage:*** Generate the second-stage design that is an $n_2 \times p$ data matrix $\mathbf{U}_2 = \left[\mathbf{D}_2, X^*_{3:p}, \ldots, X^*_{p:p}\right]$, where $\mathbf{D}_2 = \begin{pmatrix} x^{(1)}_{1:p} & \cdots & x^{(n_2)}_{1:p} \\ x^{(1)}_{2:p} & \cdots & x^{(n_2)}_{2:p} \end{pmatrix}^T$ is an optimal design from the experimental design viewpoint over the domain of the first two highest importance input factors $X_{1:p}$ and $X_{2:p}$, and $X^*_{k:p} = (x^*_{k:p}, \ldots, x^*_{k:p})^T$ is a vector that all of its n_2 values are fixed to the highest importance level value $x^*_{k:p}$ of the kth input factor $X_{k:p}$ for $k = 3, \ldots, p$. Calculate the second-stage observed output

vector via a physical experiment or exact output factor via a computer experiment, say $Y_2 = F(\mathbf{U}_2)$. Find the second-stage meta-model $\widehat{F_2}$ that is the approximate model for the relationship between the $\mathbf{D}_2$ in the second-stage design $\mathbf{U}_2$ and the corresponding second-stage observed output factor $Y_2 = F(\mathbf{U}_2)$.

- ***Third designing-modeling stage:*** Generate the third-stage design that is an $n_3 \times p$ data matrix $\mathbf{U}_3 = \left[\mathbf{D}_3, X^*_{4:p}, \ldots, X^*_{p:p}\right]$, where $\mathbf{D}_3 = \begin{pmatrix} x^{(1)}_{1:p} & \ldots & x^{(n_3)}_{1:p} \\ x^{(1)}_{2:p} & \ldots & x^{(n_3)}_{2:p} \\ x^{(1)}_{3:p} & \ldots & x^{(n_3)}_{3:p} \end{pmatrix}^T$ is an optimal design from the experimental design viewpoint over the domain of the first three highest influence inputs $X_{1:p}$, $X_{2:p}$ and $X_{3:p}$, and $X^*_{k:p} = (x^*_{k:p}, \ldots, x^*_{k:p})^T$ is a vector that all of its n_3 values are fixed to the highest influence level value $x^*_{k:p}$ of the kth input for $k = 4, \ldots, p$. Calculate the third-stage observed output vector via a physical experiment or exact output vector via a computer experiment, say $Y_3 = F(\mathbf{U}_3)$. Find the third-stage meta-model $\widehat{F_3}$ that is the approximate model for the relationship between the $\mathbf{D}_3$ in the third-stage design $\mathbf{U}_3$ and the corresponding third-stage observed output vector $Y_3 = F(\mathbf{U}_3)$.
- ***P-th designing-modeling stage*** Repeat the above systematic strategy up to the last stage as follows. Generate the pth-stage design that is an $n_p \times p$ data matrix $\mathbf{U}_p = \left[\mathbf{D}_p\right]$, where $\mathbf{D}_p = \begin{pmatrix} x^{(1)}_{1:p} & \ldots & x^{(n_p)}_{1:p} \\ \vdots & \vdots & \vdots \\ x^{(1)}_{p:p} & \ldots & x^{(n_p)}_{p:p} \end{pmatrix}^T$ is an optimal design from the experimental design viewpoint over the domain of all the p inputs $X_{1:p}, X_{2:p}, \ldots, X_{p:p}$. Calculate the pth-stage observed output vector via a physical experiment or exact output vector via a computer experiment, say $Y_p = F(\mathbf{U}_p)$. Find the pth-stage meta-model $\widehat{F_p}$ that is the approximate model for the relationship between the $\mathbf{D}_p$ in the pth-stage design $\mathbf{U}_p$ and the corresponding pth-stage observed output vector $Y_p = F(\mathbf{U}_p)$.
- ***Final Meta-Model:*** To define the overall meta-model, we use the idea of the weighted average for the coefficients of the factors in the meta-models $\widehat{F_1}, \ldots, \widehat{F_p}$. For instance as given in Fig. 1, for an experiment with three factors without interactions and the three meta-models are polynomial models as follows: $\widehat{F_1} = \beta_1 + a_{11}X_1 + a_{12}X_1^2$, $\widehat{F_2} = \beta_2 + a_{21}X_1 + a_{22}X_1^2 + b_{21}X_2 + b_{22}X_2^2$ and $\widehat{F_3} = \beta_3 + a_{31}X_1 + a_{32}X_1^2 + b_{31}X_2 + b_{32}X_2^2 + c_{31}X_3 + c_{32}X_3^2$. Therefore, the overall meta-model is the weighted average that is given as follows:

$$\widehat{F} = \frac{1}{3}\sum_{k=1}^{3}\beta_k + \left(\sum_{k=1}^{3}\frac{a_{k1}}{3}\right)X_1 + \left(\sum_{k=1}^{3}\frac{a_{k2}}{3}\right)X_1^2 + \left(\sum_{k=2}^{3}\frac{b_{k1}}{2}\right)X_2 + \left(\sum_{k=2}^{2}\frac{b_{k2}}{2}\right)X_2^2 + c_{31}X_3 + c_{32}X_3^2.$$

Now comes to mind the following logical question: *How to select the optimal design (dataset) from the experimental design viewpoint for each stage over the domain of the input factors in each stage*? An efficient way for selecting optimal representative training datasets for the new proposed SeqST is to make use of the techniques of experimental design approach. The optimality selection of experimental points (design) from an experimental region that provides valuable information about a given experiment is the most significant

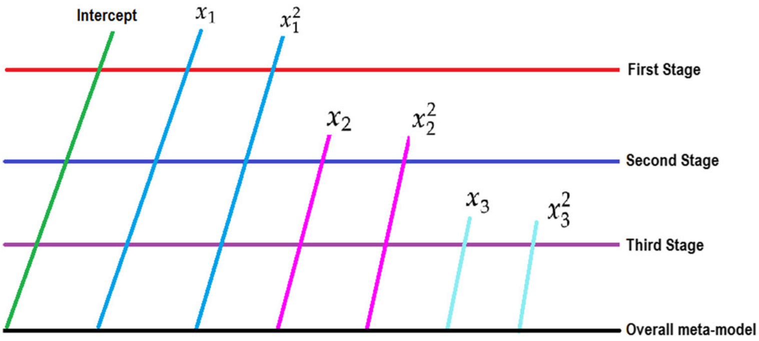

Fig. 1 The main idea of the weighted average to get the final meta-model

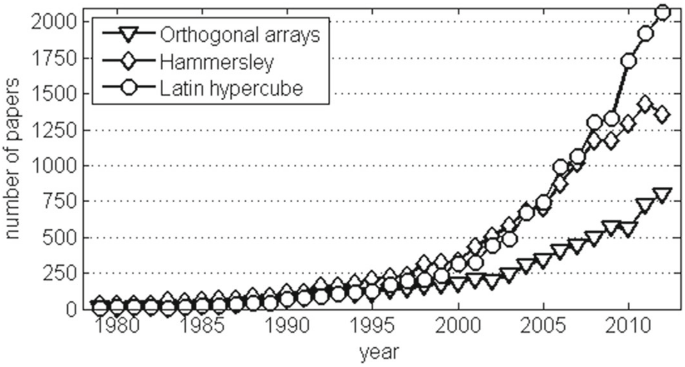

Fig. 2 Number of papers published per year. Data obtained from the Google Scholar database in the week of March 4, 2013 (cf. Fig. 1a in Viana 2013)

hard problem investigators may face, especially when there is no prior information about the model structure between the inputs and the corresponding outputs. An intuitive idea to overcome the mentioned problem is to scatter the representative training points in an intelligent manner to cover the experimental region well, which is called a space-filling design (cf. Elsawah 2022c). Among strategies coined for computer experiments, Latin hypercube designs (LHDs) (Mckay et al. 1979; Iman and Conover 1980) have become very popular. Other strategies include orthogonal arrays (Owen 1992), and Hammersley designs (Diwekar and kalagnanam 1997; Hammersley 1960). To illustrate their popularity, Fig. 1a in Viana (2013) (cf. Fig. 2) shows an approximate number of publications that referred to at least one of these three techniques. An LHD spreads its representative training points everywhere in the region with as few gaps or holes as possible (cf. Fig. 3), and thus, it gives a good representation of the experimental region with even fewer points. LHDs play an important role in computer simulation (cf. Husslage et al. 2011; Fang et al. 2006; Elsawah and Gong 2023). Therefore, LHDs is used in this study. It is pertinent to point out that the new proposed SeqST can be carried out utilizing uniform designs, which are a class of optimal space-filling designs that are currently extensively used in a variety of practical applications (cf. Elsawah and Vishwakarma 2022).

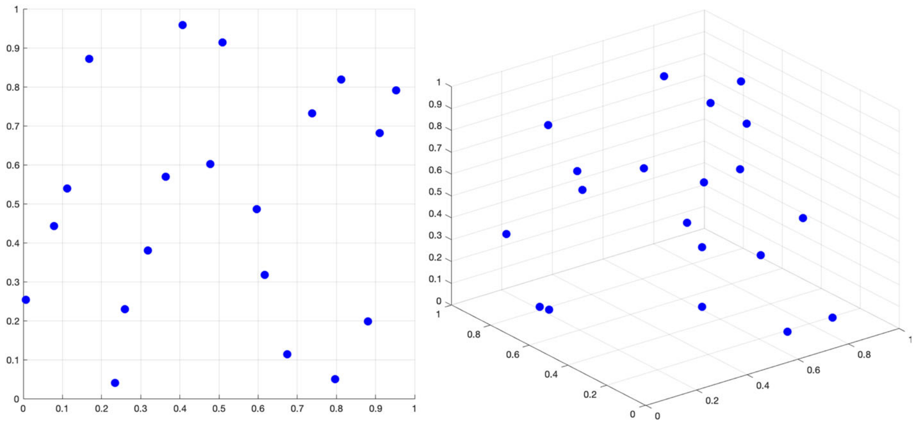

Fig. 3 Latin hypercube 20 points in two dimensions and three dimensions

3 On the importance of the input factors and their points

The following logical question comes to mind after reading the preparation stage of the new proposed SeqST: *If there is no prior information, how to determine the order of the importance of the factors and the points with the highest influence for each factor*? This section tries to provide an answer to this significant question for computer experiments. Consider a computer experiment with p independent input factors $X_i \in [LB_i, UB_i],\ 1 \leq i \leq p$ and x_k^* is the point with the highest influence for the kth factor $X_k,\ 1 \leq k \leq p$. From physics point of view, the points x_k^* with the highest influence can be defined as the Mass Centers (MCs). The MC is a point that causes a rigid body to maintain its equilibrium state. Within a solid Q with volume V, if the mass distribution is continuous with density ρ, the integral of the weighted position coordinates of the points connected to the center of mass R can be expressed as follows:

$$\iiint_Q \rho(r)(r - R)\,\mathrm{d}V = 0, \tag{1}$$

where r is the vector representing the position of a point with respect to a fixed origin and the solution of coordinate R is given as follows:

$$R = \frac{1}{M}\iiint_Q \rho(r) r\,\mathrm{d}V, \tag{2}$$

where M is the total mass of the solid. For further details, the reader may refer to Mark (2009). If the body is formed by a function from mathematics viewpoint, its volume has a uniform density distributed state with a constant $\rho(r)$. Therefore, for a function with p factors $F(X_1, X_2, \ldots, X_p)$, (2) can be rewritten as follows:

$$R = \frac{1}{M}\underbrace{\int \cdots \int \cdots \int}_{\text{p integrals}} F(X_1, X_2, \ldots, X_p)\,\mathrm{d}V. \tag{3}$$

The point x_k^* with the highest influence for the factor X_k is defined as the point that divides the function into two parts with same mass, $M_L = M_R$ (cf. Fig. 4). Therefore, from (3), we

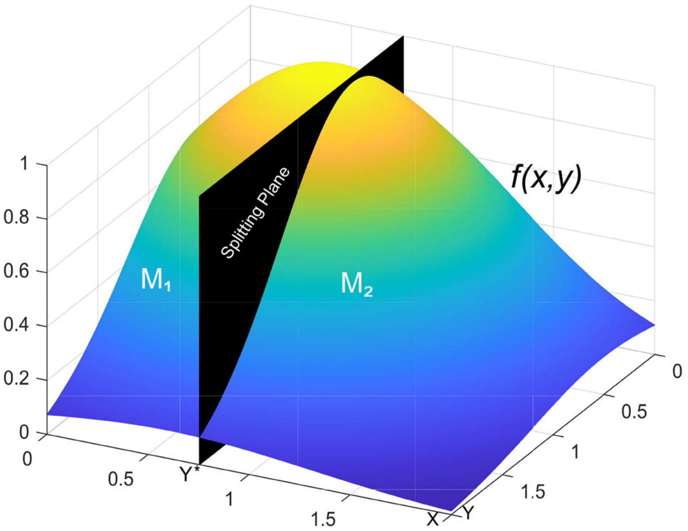

Fig. 4 The mass centers of a function

get

$$\frac{1}{R}\underbrace{\int\cdots\int\cdots\int}_{\text{p integrals}} F(X_1, X_2, \ldots, X_p)\,\mathrm{d}V_L = \frac{1}{R}\underbrace{\int\cdots\int\cdots\int}_{\text{p integrals}} F(X_1, X_2, \ldots, X_p)\,\mathrm{d}V_R. \tag{4}$$

From (4), the point x_k^* with the highest influence for the factor X_k is the solution of the following equation:

$$\begin{aligned} &\int_{LB_p}^{UB_p}\cdots\int_{LB_k}^{x_k^*}\cdots\int_{LB_1}^{UB_1} F(X_1, \ldots, X_k, \ldots, X_p)\ \mathrm{d}X_1 \ldots \mathrm{d}X_k \ldots \mathrm{d}X_p \\ &\quad = \int_{LB_p}^{UB_p}\cdots\int_{x_k^*}^{UB_k}\cdots\int_{LB_1}^{UB_1} F(X_1, \ldots, X_k, \ldots, X_p)\ \mathrm{d}X_1 \ldots \mathrm{d}X_k \ldots \mathrm{d}X_p. \end{aligned} \tag{5}$$

Using the calculated points $x_k^*,\ 1 \le k \le p$ with the highest impacts, the importance of the factor X_k can be measured by its corresponding area as follows:

$$A(X_k) = \left|\int_{LB_k}^{UB_k} F(x_1^*, x_2^*, \ldots, X_k, \ldots, x_p^*)\,\mathrm{d}X_k\right|. \tag{6}$$

The p areas $A(x_k),\ 1 \le k \le p$ need to be calculated and sorted in a decreasing order as follows:

$$A(X_{1:p}) > A(X_{2:p}) > \cdots > A(X_{p:p}),$$

where $X_{1:p}$ is the input with the highest importance and $X_{p:p}$ is the input with the lowest importance.

4 An illustrative example

The above-mentioned steps and discussions in Sects. 2 and 3 are used and explained using LHDs, polynomial regression models, and the following computer experiment:

$$Y=F(X_1, X_2, X_3)=200+5X_1^2+100X_1+\frac{1}{25}X_2^2+50X_2+\frac{1}{175}X_3^2+X_3,\ 0\le X_i\le 1,\ 1\le i\le 3.$$

Based on (5), the points with the highest impacts for the factors X_1, X_2, and X_3 are calculated as follows $x_1^* = 0.5470$, $x_2^* = 0.5225$, and $x_3^* = 0.5005$, respectively. Based on (6), the corresponding areas for the factors X_1, X_2, and X_3 are calculated as follows $A(X_1) = 51.67$, $A(X_2) = 25.01$, and $A(X_3) = 0.5019$, respectively. Therefore, the order of the importance of the input factors is given as follows $X_1 \ggg X_2 \ggg X_3$. Table 1 gives LHDs with 11, 16, and 20 points for the first, second, and third stages, respectively, and their corresponding outputs. From Table 1 and the proposed SeqST in Sect. 2, we get

- The first meta-model $\widehat{F_1}$ gives the following relationship between the LHD $\mathbf{D}_1 = [X_1]$ and the corresponding output $Y_1 = F(\mathbf{U}_1)$:
$$\widehat{F_1} = 278.3730 + 30.4786X_1.$$
- The second meta-model $\widehat{F_2}$ gives the following relationship between the LHD $\mathbf{D}_2 = [X_1\ X_2]$ and the corresponding output $Y_2 = F(\mathbf{U}_2)$:
$$\widehat{F_2} = 278.1910 + 31.5193X_1 + 1.6511X_1^2 + 15.1802X_2 + 0.0123X_2^2.$$
- The third meta-model $\widehat{F_3}$ gives the following relationship between the LHD $\mathbf{D}_3 = [X_1\ X_2\ X_3]$ and the corresponding output $Y_3 = F(\mathbf{U}_3)$:
$$\begin{aligned}\widehat{F_3} = {} & 275.0492 + 29.9669X_1 + 1.5248X_1^2 + 15.8355X_2 \\ & +0.0124X_2^2 + 0.2970X_3 + 0.0018X_3^2.\end{aligned}$$

Therefore, the overall meta-model is the weighted average that is given as follows:

$$\begin{aligned}\widehat{F}_{SeqST} = {} & 277.2204 + 30.6549X_1 + 15.5078X_2 + 0.2970X_3 + 1.5880X_1^2 \\ & +0.0123X_2^2 + 0.0018X_3^2.\end{aligned}$$

To test the performance of this meta-model, the SinST is used to find another meta-model using an LHD $\mathbf{U}$ with the same number of points in the three stages of the SeqST, i.e., $n = n_1 + n_2 + n_3 = 11 + 16 + 20 = 47$. Table 2 gives an LHD $\mathbf{U} = [X_1\ X_2\ X_3]$ and the corresponding output $Y = F(\mathbf{U})$. From Table 2, the meta-model that describes the relationship between $\mathbf{U}$ and $Y = F(\mathbf{U})$ is given as follows:

$$\begin{aligned}\widehat{F}_{SinST} = {} & 277.2204 + 28.9878X_1 + 14.4962X_2 + 0.2900X_3 \\ & +1.4974X_1^2 + 0.0120X_2^2 + 0.0017X_3^2.\end{aligned}$$

Figure 5 gives all the 47 values of $F(\mathbf{U}_{\text{test}})$, $\widehat{F}_{SeqST}(\mathbf{U}_{\text{test}})$ and $\widehat{F}_{SinST}(\mathbf{U}_{\text{test}})$ and the absolute differences between each two of them using an LHD $\mathbf{U}_{\text{test}}$ with 47 points as a testing dataset. The results show that the values of $\widehat{F}_{SeqST}(\mathbf{U}_{\text{test}})$ are closer to $F(\mathbf{U}_{\text{test}})$ than the values of $\widehat{F}_{SinST}(\mathbf{U}_{\text{test}})$. Moreover, the mean square error (MSE), $MSE = \frac{1}{n}\sum_{i=1}^{n}(F_i - \widehat{F}_i)^2$, of these two meta-models are given as follows:

$$MSE_{SeqST} = 1.0875 \times 10^3 < MSE_{SinST} = 1.6838 \times 10^3.$$

Therefore, the SeqST is much better than the SinST.

Table 1 The three-stage designs and their corresponding observed outputs for the SeqST for the illustrative example

Third stage					Second stage					First stage				
$\mathbf{U}_3 = [\mathbf{D}_3]$					$\mathbf{U}_2 = [\mathbf{D}_2\ X_3]$					$\mathbf{U}_1 = [\mathbf{D}_1\ X_2\ X_3]$				
n_3	X_1	X_2	X_3	$Y_3 = F(\mathbf{U}_3)$	n_2	X_1	X_2	X_3	$Y_2 = F(\mathbf{U}_2)$	n_1	X_1	X_2	X_3	$Y_1 = F(\mathbf{U}_1)$
1	0.9340	0.0547	0.0418	300.5415	1	0.8594	0.8804	0.5005	334.1884	1	0.2196	0.5225	0.5005	248.8382
2	0.0888	0.9565	0.2782	257.0558	2	0.3719	0.0514	0.5005	240.9501	2	0.8086	0.5225	0.5005	310.7706
3	0.6302	0.6285	0.5479	297.0026	3	0.7528	0.8106	0.5005	319.1757	3	0.1221	0.5225	0.5005	238.9224
4	0.7409	0.9227	0.1660	323.1693	4	0.5775	0.8106	0.5005	275.6280	4	0.3132	0.5225	0.5005	258.4470
5	0.1094	0.1802	0.4338	220.4423	5	0.3064	0.5310	0.5005	258.1724	5	0.9945	0.5225	0.5005	331.0326
6	0.5713	0.4329	0.9739	281.3934	6	0.4501	0.6934	0.5005	281.2128	6	0.5009	0.5225	0.5005	277.9786
7	0.9901	0.7502	0.7748	342.2193	7	0.7417	0.9438	0.5005	324.6537	7	0.7097	0.5225	0.5005	300.1246
8	0.3342	0.8932	0.3859	279.0551	8	0.1933	0.6799	0.5005	254.0346	8	0.5821	0.5225	0.5005	286.5436
9	0.4647	0.4554	0.4952	270.8274	9	0.3896	0.8654	0.5005	283.5216	9	0.8991	0.5225	0.5005	320.5928
10	0.3999	0.6532	0.7330	274.2021	10	0.0652	0.1438	0.5005	214.2358	10	0.3942	0.5225	0.5005	266.8388
11	0.2157	0.7177	0.6583	258.3661	11	0.0464	0.5705	0.5005	233.6837	11	0.0329	0.5225	0.5005	229.9309
12	0.0202	0.1320	0.3108	208.9374	12	0.6622	0.2095	0.5005	279.3887					
13	0.7576	0.5045	0.8243	304.6907	13	0.1432	0.4065	0.5005	235.2541					
14	0.6921	0.8212	0.8800	313.5756	14	0.9363	0.2896	0.5005	321.9976					
15	0.5036	0.0207	0.0896	252.7510	15	0.9497	0.4910	0.5005	324.5390					
16	0.1532	0.3124	0.6463	231.7120	16	0.5021	0.0811	0.5005	256.0279					
17	0.8139	0.2568	0.1382	297.6847										
18	0.4180	0.5937	0.2289	272.6018										
19	0.2926	0.3928	0.5608	249.8930										
20	0.8930	0.2287	0.9397	305.6684										

Table 2 The single-stage design and its corresponding observed outputs for the SinST for the illustrative example

n	X_1	X_2	X_3	$Y = F(\mathbf{U})$	n	X_1	X_2	X_3	$Y = F(\mathbf{U})$
1	0.4700	0.4770	0.7960	272.7220	25	0.4610	0.3460	0.6810	265.1510
2	0.9060	0.7380	0.5440	332.1120	26	0.7850	0.0460	0.7300	284.6150
3	0.2400	0.7770	0.5740	263.7330	27	0.7510	0.2950	0.4220	293.0740
4	0.7070	0.2170	0.4490	284.4700	28	0.9370	0.0280	0.2120	299.7660
5	0.9350	0.1410	0.3190	305.2530	29	0.3020	0.1550	0.5200	238.9340
6	0.4370	0.1830	0.7590	254.5380	30	0.8240	0.9560	0.3570	333.9830
7	0.9780	0.3280	0.1300	319.0950	31	0.8770	0.1210	0.0070	297.6070
8	0.4920	0.8420	0.9280	293.4340	32	0.6010	0.6050	0.8430	292.9440
9	0.9820	0.5890	0.2510	332.7590	33	0.1420	0.4220	0.8890	236.2810
10	0.3580	0.6600	0.3640	269.7720	34	0.2260	0.2530	0.2140	235.7230
11	0.8460	0.3760	0.8130	307.7530	35	0.0420	0.4670	0.2930	227.8860
12	0.2700	0.1930	0.1150	237.1770	36	0.5480	0.6300	0.2720	288.1210
13	0.6770	0.7020	0.0610	305.1390	37	0.3970	0.0100	0.3900	241.3430
14	0.2010	0.4350	0.1510	242.2310	38	0.4230	0.2730	0.0900	256.9080
15	0.3790	0.9740	0.4930	287.8450	39	0.0750	0.7980	0.7190	248.1550
16	0.1190	0.5060	0.5880	237.8710	40	0.6440	0.6400	0.9910	299.4560
17	0.7380	0.6960	0.6400	312.0270	41	0.8550	0.5540	0.3380	317.2380
18	0.5130	0.9830	0.9460	302.6880	42	0.2820	0.8220	0.9020	270.6050
19	0.1630	0.5530	0.1860	244.2760	43	0.0060	0.0990	0.6770	206.2350
20	0.1910	0.9280	0.4810	266.1780	44	0.3270	0.3170	0.6090	249.7410
21	0.5720	0.5110	0.6350	285.0590	45	0.5950	0.4000	0.9600	282.2200
22	0.8070	0.7630	0.8630	323.0040	46	0.6380	0.8620	0.0280	308.9410
23	0.0860	0.8830	0.0810	252.8990	47	0.6910	0.9140	0.4370	317.6490
24	0.0480	0.0770	0.7860	209.4360					

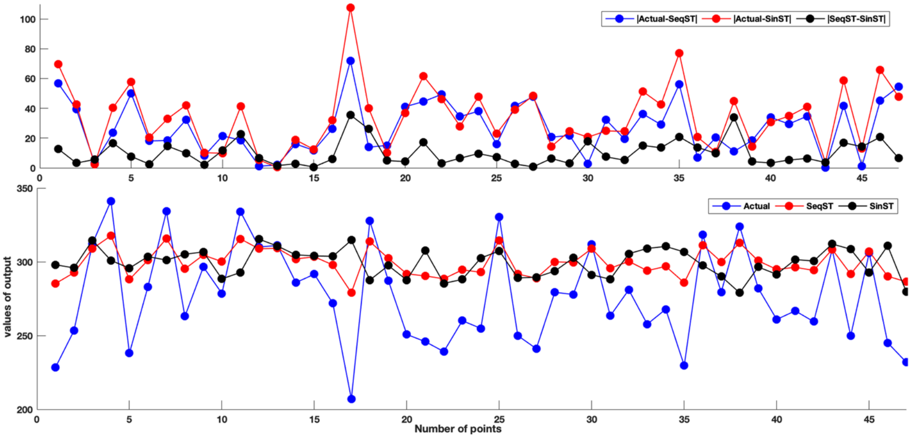

Fig. 5 All the 47 values of $F(\mathbf{U}_{\text{test}})$, $\widehat{F}_{SeqST}(\mathbf{U}_{\text{test}})$ and $\widehat{F}_{SinST}(\mathbf{U}_{\text{test}})$ (down) and the absolute differences between each two of them (up) using an LHD $\mathbf{U}_{\text{test}}$ with 47 points as a testing dataset for the illustrative example

5 The performance assessment of the new proposed SeqST

To evaluate the performance of our proposed methodology, we consider the following four examples, two linear models and two non-linear models. The first linear model is the so-called the pullulan production model. Although pullulan has been produced commercially since 1978, the production mechanism on the genetic level is still far from being fully understood. As a result, only empirical models can be built to optimize pullulan production. One of these models is derived by Goksungur et al. (2005) as follows:

$$\begin{aligned} Y_1 = &-29.851 + 1.189X_1 + 0.057X_2 + 5.086X_3 - 0.011X_1^2 - 0.0000607X_2^2 - 1.3633X_3^2 \\ &-0.000296X_1X_2 + 0.0263X_1X_3. \end{aligned}$$

This model predicts the final concentration of pullulan (g/L) as a function of the initial substrate concentration (X_1), the speed of agitation (X_2), and the airflow rate (X_3). The ranges of variation of the independent variables are $X_1 \in [30\ 70]$ g/L, $X_2 \in [200\ 600]$ rpm, and $X_3 \in [1\ 3]$ vvm. The range of variation of the dependent variable Y_1 is [4.96 17]. The second linear model is the so-called the Goldprice model that has been studied by Andre et al. (2000) and Ranjan et al. (2008). The Goldprice function is given by

$$\begin{aligned} Y_2 = &\left[1 + (X_1 + X_2 + 1)^2 \left(19 - 14X_1 + 3X_1^2 - 14X_2 + 6X_1X_2 + 3X_2^2\right)\right] \\ &\times \left[30 + (2X_1 - 2X_2)^2 \left(18 - 32X_1 + 12X_1^2 + 48X_2 - 36X_1X_2 + 27X_2^2\right)\right], \end{aligned}$$

where the two input factors X_1 and X_2 are defined on the domain $[-2\ 2] \times [-2\ 2]$.

The non-linear model is an equation selected for its very different topology and non-linearity compared to the first two models. The first non-linear model is given as follows:

$$Y_3 = \frac{\ln(X_1)(\sin X_2 + 4)}{\exp(X_3)} + \ln(X_1)\exp(X_3),$$

where the ranges of variation of the independent variables are $X_1 \in [0.1\ 10]$, $X_2 \in [-\pi/2\ \pi/2]$, and $X_3 \in [0\ 1]$ leading to a variation of the dependent variable Y in the range of $[-13.82\ 13.82]$. The second non-linear model is given as follows:

$$Y_4 = \exp(X_1) + \sin(X_2) + X_3^7,$$

where the range of variation of the independent variables is [0 1].

A comparison study between the mean squared error (MSE), $MSE = \frac{1}{n}\sum_{i=1}^{n}(F_i - \widehat{F_i})^2$, and mean absolute error (MAE), $MAE = \frac{1}{n}\sum_{i=1}^{n}|F_i - \widehat{F_i}|$, of the meta-models using the new proposed SeqST and the classical SinST is given based on the above-mentioned four models using the LHDs as training and testing datasets, the polynomial models as the fitting models, and the medians of the ranges of the input factors as the points with the highest impacts. To have a fair comparison study between the SeqST and SinST, the number of representative training points for SinST is selected to be equal to the total number of representative training points in all the p stages of the SeqST, i.e., $n = n_1 + n_2 + \cdots + n_p$. Since the representative training datasets and the representative testing datasets (LHDs) are not deterministic for a given n, the minimum, mean, median, and 95% confidence interval (95%CI) of the MSEs and MAEs of the approximate meta-models of the above-mentioned four models using the SeqST and SinST based on about 5000 different randomly generated representative training datasets and representative testing datasets are given in Table 3 to investigate the behavior of the SeqST for any randomly generated representative datasets. From Table 3, we get the following:

Table 3 The simulation results for the performance of the SeqST and SinST using the above-mentioned models Y_1, Y_2, Y_3 and Y_4 via 5000 repetitions

Method	Criterion	Minimum	Mean	Median	95% CI
The first linear model: The pullulan production model (Y_1)					
SeqST	MSE	20.091	23.101	23.049	[23.074 23.128]
	MAE	3.8911	4.0641	4.0626	[4.0626 4.0656]
SinST	MSE	22.848	25.682	25.648	[25.658 25.706]
	MAE	4.1542	4.3504	4.3497	[4.3489 4.3519]
The second linear model: The Goldprice model (Y_2)					
SeqST	MSE	7.7708E+8	5.2088E+9	5.0384E+9	[5.1527E+9 5.2648E+9]
	MAE	14290	30862	28500	[30621 31104]
SinST	MSE	8.0513E+8	5.2943E+9	5.0967E+9	[5.2401E+9 5.3484E+9]
	MAE	15360	35140	31708	[34853 35428]
The first non-linear model (Y_3)					
SeqST	MSE	1.2137	2.2219	1.8711	[2.0348 2.409]
	MAE	0.85987	1.1416	1.0907	[1.1318 1.1514]
SinST	MSE	2.55	4.1431	4.0419	[4.1264 4.1598]
	MAE	1.2682	1.6338	1.6167	[1.6303 1.6373]
The second non-linear model (Y_4)					
SeqST	MSE	0.22297	0.31334	0.29591	[0.30076 0.32591]
	MAE	0.39099	0.46239	0.45811	[0.45988 0.4649]
SinST	MSE	0.3105	0.43789	0.43689	[0.4369 0.43888]
	MAE	0.45426	0.55088	0.55035	[0.55018 0.55158]

- The new proposed SeqST is better than the classical SinST for all the four models, where the values of the MSE and MAE via the SeqST are smaller than their values via the SinST. The SeqST is better than the SinST for 5000 different training and testing datasets, where the minimum, mean, and median of about 5000 MSE and MAE values via the SeqST are less than their values via the SinST for all the cases.
- The gaps among the impacts of the input factors for $Y_3 >$ (i.e., greater than) the gaps among the impacts of the input factors for $Y_4 >$ the gaps among the impacts of the input factors for $Y_1 >$ the gaps among the impacts of the input factors for Y_2. The performance of the SeqST for $Y_3 \succeq$ (i.e., better than) the performance of the SeqST for $Y_4 \succeq$ the performance of the SeqST for $Y_1 \succeq$ the performance of the SeqST for Y_2, where the percentage differences between the minimum, mean, and median of the MSEs (and MAEs) for the SeqST and SinST for $Y_3 >$ that for $Y_4 >$ that for $Y_1 >$ that for Y_2. That is, when there are significant gaps among the impacts of the input factors, the accuracy of the SeqST increases.

6 Further interesting investigation for the performance of the SeqST

After the above-mentioned results come to mind the following new logical questions: *What is the effect of the order of the importance of the input factors on the accuracy of the new proposed SeqST? What is the effect of the gaps among the importance of the input factors on* *the accuracy of the new proposed SeqST? What is the effect of the total number of points on the accuracy of the new proposed SeqST? What is the effect of the number of points in each stage on the accuracy of the new proposed SeqST?* The answers of these questions provide benchmarks for the optimality use of the new proposed SeqST. This section tries to answer these questions and other interesting questions based on computer experiments.

Let the following non-linear model:

$$Y_5 = -e^{-(X_1+0.5)^2} - 2e^{-(X_2-0.5)^2} - 4e^{-(X_3+3)^2},\ 0 \le X_i \le 1,\ 1 \le i \le 3.$$

Figure 6 investigates the importance of the three input factors for the model Y_5. From Fig. 6 and based on the area under each curve, we get that $X_2 \ggg X_1 \ggg X_3$ is the order of the importance of Y_5. To check the effect of the number of points in each stage and the order of the importance of the input factors on the accuracy of the SeqST, different numbers of points in each stage are used as follows: $10 \le n_i \le 100,\ 1 \le i \le 3$ and $n_1+n_2+n_3 = 120$. Figures 7, 8, and 9 give the MSE values of the SeqST for the model Y_5 using different number of points in each stage based on the following three different order of importance: $X_2 \ggg X_1 \ggg X_3$ (correct order), $X_1 \ggg X_2 \ggg X_3$ (wrong order), and $X_3 \ggg X_2 \ggg X_1$ (wrong order), respectively. Figure 10 gives a comparison study between the SeqST and SinST based on different number of points in each stage from the three stages of Y_5, where the number of points in the SinST n is equal to the number of points in the three stages of the SeqST, i.e., $n = n_1 + n_2 + n_3$. From Figs. 7, 8, 9 and 10, we conclude that:

- The MSE values using the correct order of the importance are less than the MSE values using the wrong order of the importance for any number of points in each stage, where the ranges of MSE values are about (0.06 0.073), (0.24 0.65), and (0.43 0.63) for $X_2 \ggg X_1 \ggg X_3$ (correct order), $X_1 \ggg X_2 \ggg X_3$ (wrong order), and $X_3 \ggg X_2 \ggg X_1$ (wrong order), respectively. Therefore, *it is recommended to carefully check the order of the importance before using the new proposed SeqST.*
- The new proposed SeqST is better than the classical SinST for any number of points, where the MSE values for the SeqST are less than the MSE values for the SinST. However, the SeqST is much better than the SinST for a small number of points (cf. Fig. 10). Therefore, *it is recommended to use the new proposed SeqST for small number of points (experiments with a few trials).*

Moreover, from the discussions about the models Y_1, Y_2, Y_3, and Y_4 in Sect. 5, it is observed that: *When there are significant gaps among the impacts of the input factors, the accuracy of the new proposed SeqST increases*. The following discussion tries to give more investigations for this interesting observation using two different types of gaps among the impacts of the input factors. The first type is the power gap that is investigated using the following model:

$$Y_6 = X_1^{\alpha_1} + X_2^{\alpha_2} + X_3^{\alpha_3},\ 0 \le X_i \le 1,\ 1 \le \alpha_i \le 8,\ 1 \le i \le 3,\ 3 \le \alpha_1 + \alpha_2 + \alpha_3 \le 10.$$

The second type is the coefficient gap that is investigated using the following model:

$$Y_7 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,\ 0 \le X_i \le 1,\ 1 \le \beta_i \le 18,\ 1 \le i \le 3,\ 3 \le \beta_1 + \beta_2 + \beta_3 \le 20.$$

Figures 11 and 12 give the differences of the medians of the MSE values using the new proposed SeqST and the medians of the MSE values using the classical SinST based on about 5000 different randomly generated representative training datasets and representative testing datasets for different powers and coefficients of the models Y_6 and Y_7, respectively. The order is taken here as: $X_1 \ggg X_2 \ggg X_3$. That is, the correct power that is consistent

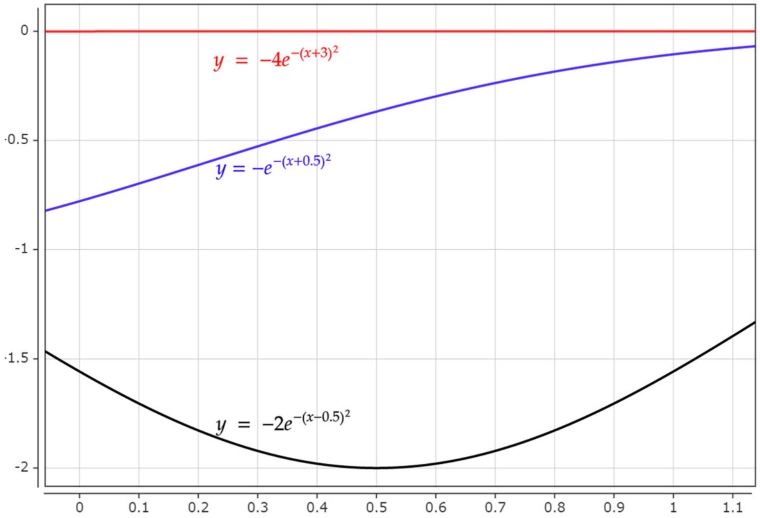

Fig. 6 The importance of the inputs for the model Y_5

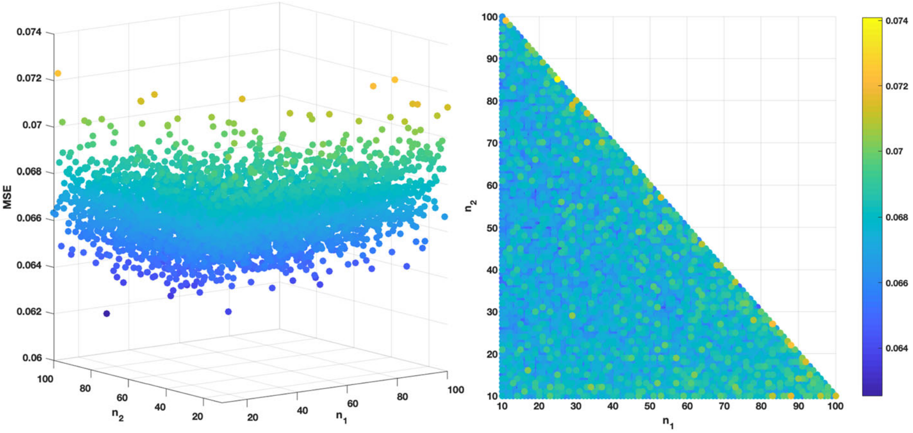

Fig. 7 The MSE for order $X_2 \ggg X_1 \ggg X_3$ (correct order) for the model Y_5

with this order is $\alpha_1 < \alpha_2 < \alpha_3$; however, the correct coefficient that is consistent with this order is $\beta_1 > \beta_2 > \beta_3$. From Figs. 11 and 12, we get

- When there are big gaps among the impacts of the input factors, the performance of the new proposed SeqST is much better than the performance of the classical SinST. Keep in mind that $0 \leq X_i \leq 1$, i.e., when there are big gaps among the powers, α_1, α_2, and α_3, we have small gaps among the impacts of the input factors, X_1, X_2, and X_3, and vice versa. However, when there are big gaps among the coefficients, β_1, β_2, and β_3, we have big gaps among the the impacts of the input factors, X_1, X_2, and X_3, and vice versa. Therefore, *it is recommended to use the new proposed SeqST for experiments with large gaps among the impacts of their input factors*.
- For small powers α_1 and α_2 and large power α_3 (i.e., correct order of the importance), the performance of the new proposed SeqST is much better than its performance for

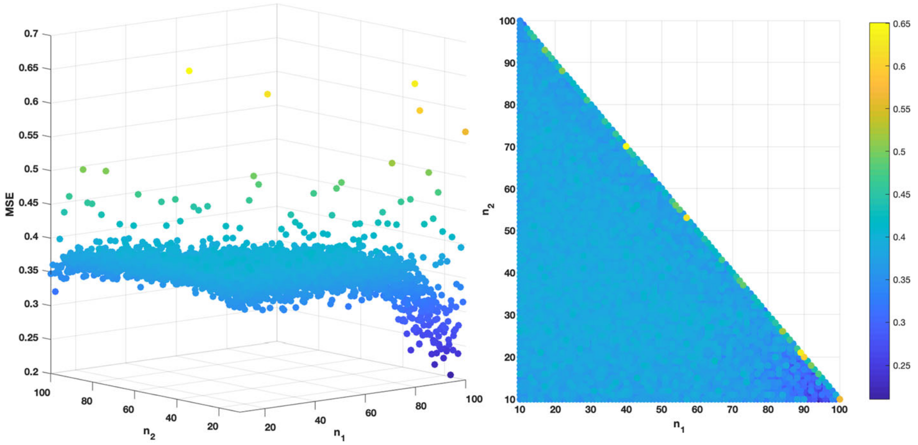

Fig. 8 The MSE for order $X_1 \ggg X_2 \ggg X_3$ (wrong order) for the model Y_5

large power α_3 (i.e., wrong order of the importance). For large coefficients β_1 and β_2 and small coefficient β_3 (i.e., correct order of the importance), the performance of the SeqST is much better than its performance for large coefficient β_3 (i.e., wrong order of the importance). Therefore, we get the same conclusion that is mentioned above: *It is recommended to carefully check the importance order before using the new proposed SeqST.*

To provide a more investigation to the effect of the number of points in each stage on the accuracy of the new proposed SeqST, let the following model:

$$Y_8 = X_1^4 + \frac{1}{2}X_2^4 + \frac{1}{3}X_3^4,\ 0 \le X_i \le 1,\ 1 \le i \le 3.$$

Figure 13 investigates the importance of the three input factors for the models Y_8. From Fig. 13 and based on the area under each curve, we get that $X_1 \ggg X_2 \ggg X_3$ is the order of the importance of Y_8. Table 4 gives the MSE values and MAE values for Y_8 based on the correct order of the importance and different number of training points in each stage. Moreover, Table 4 gives the MSE values and MAE values for the above-mentioned Y_4 and Y_5. From Table 4, we conclude that: $n_1 > n_2 > n_3$ is the best selection of the number of the training points in the three stages. Therefore, *it is recommended to use the new proposed SeqST with a descending order of the numbers of training points in its stages.*

7 Conclusion and future work

This paper gives a new sequential stage technique (SeqST) for designing and modeling experiments when the input factors are not equally important. In the new proposed SeqST, the input factors are added to the process and modeled sequentially according to their importance, one input factor is added at each stage, while each remaining input factor keeps fixed at a given point that has the highest influence. A comparison study between the new proposed SeqST and the classical single-stage technique (SinST) is investigated. The effects of: the order of the importance of the input factors, the number of the training points in each stage, the total number of the training points, and the gaps among the influences of the input factors, on

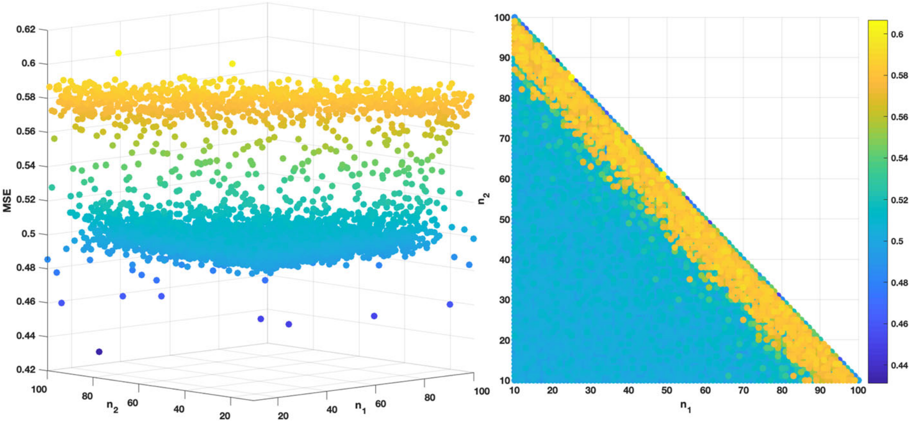

Fig. 9 The MSE for order $X_3 \ggg X_2 \ggg X_1$ (wrong order) for the model Y_5

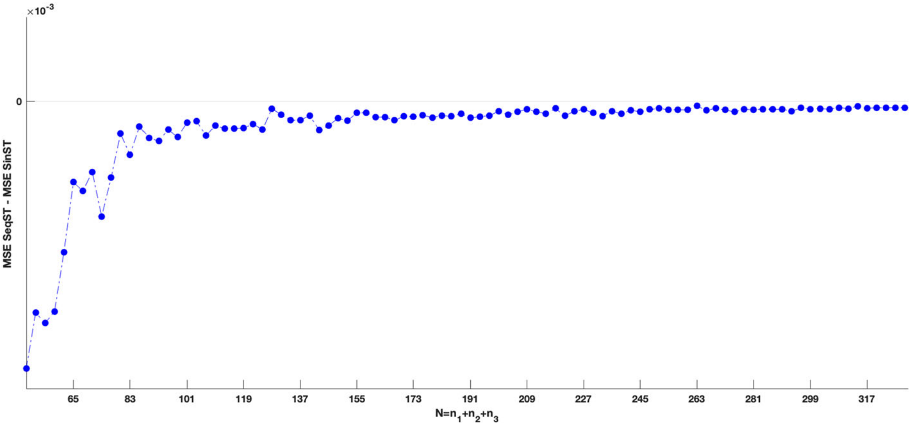

Fig. 10 The MSESeqST–MSESinST for the model Y_5

the performance of the new proposed SeqST are investigated. This study gives a benchmark that guide experimenters to effectively designing and modeling their experiments. The main results show that:

- The performance of the new proposed SeqST is better than the performance of the classical SinST under different experimental conditions and scenarios.
- The deviation between the performance of the new proposed SeqST and the classical SinST for small number of training points is larger than that when there are a large number of training points.
- The deviation between the performance of the new proposed SeqST and the SinST for experiments with large gaps among the impacts of their factors is larger than that when there are small gaps among the impacts of their factors.
- The new proposed SeqST has a good performance using the correct order of the importance of the input factors.
- The new proposed SeqST has a good performance using a descending order of the numbers of the training points in its stages.

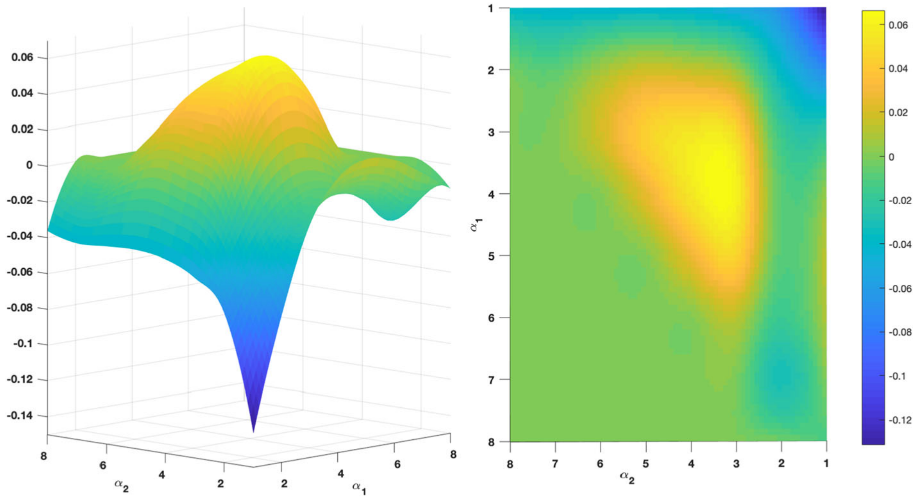

Fig. 11 The Median MSE SeqST–Median MSE SinST for the model Y_6

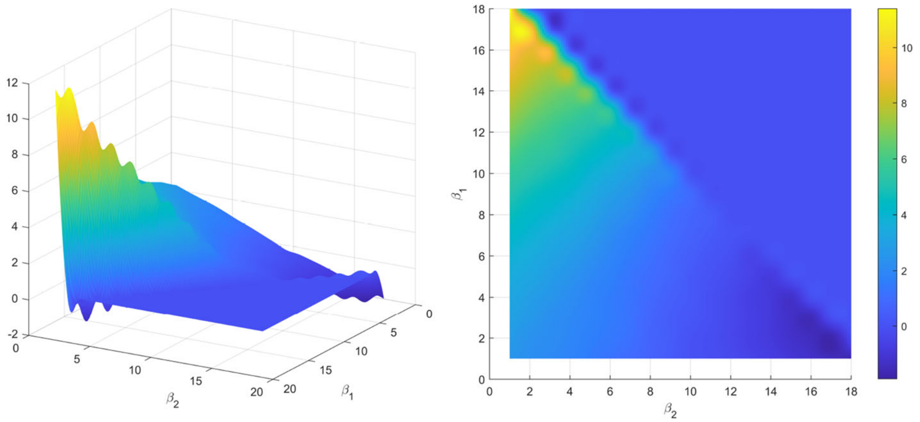

Fig. 12 The Median MSE SeqST–Median MSE SinST for the model Y_7

Therefore, we conclude that the new proposed SeqST is highly recommended to be used with the correct order of the importance of the input factors using a descending order of the training points in its stages for experiments with a few trials and/or large gaps between the importance of their factors.

During this work, the following interesting new ideas for future work have been arisen. The first author is working on them, and some theoretical and simulation results are obtained. However, more time and effort are needed to crystallize them in high-quality research papers with significant results.

- This paper is a good first stone toward more future work in this regard. For instance, there is a significant need to theoretically study the behavior of the new proposed SeqST more deeply. In this study, the LHDs are used as training and testing datasets and the polynomial model is used as the fitting model. The logical questions are that: *What is the effect of the type of training and testing datasets on the performance of the new proposed SeqST?*

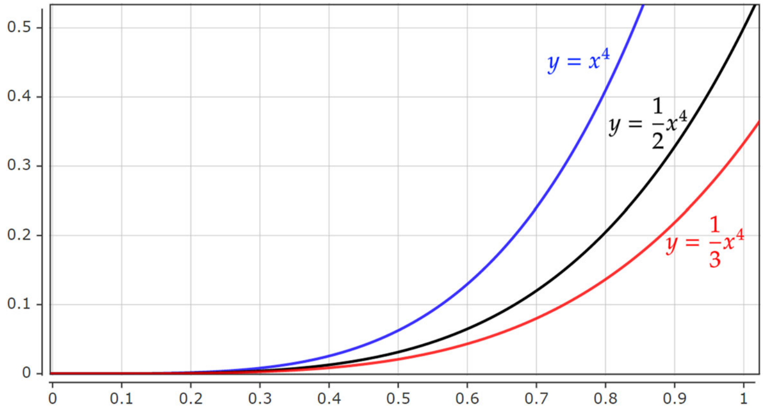

Fig. 13 The importance of the inputs for the model Y_8

Table 4 The MSE and MAE for different number of points

n_1	n_2	n_3	MSE	MAE
$Y_8 = X_1^4 + \frac{1}{2}X_2^4 + \frac{1}{3}X_3^4,\ 0 \le X_i \le 1,\ 1 \le i \le 3$				
30	50	70	0.2349	0.2365
50	50	50	0.2318	0.2324
70	50	30	0.2239	0.2240
$Y_4 = \exp(X_1) + \sin(X_2) + X_3^7,\ 0 \le X_i \le 1,\ 1 \le i \le 3$				
30	50	70	6.1417×10^7	6.4111×10^7
50	50	50	6.0332×10^7	6.2526×10^7
70	50	30	5.9626×10^7	5.9955×10^7
$Y_5 = -e^{-(X_1+0.5)^2} - 2e^{-(X_2-0.5)^2} - 4e^{-(X_3+3)^2},\ 0 \le X_i \le 1,\ 1 \le i \le 3$				
30	50	70	15.9514	16.0475
50	50	50	15.6240	15.8654
70	50	30	15.2965	15.7727

What is the effect of the type of fitting model on the performance of the new proposed SeqST? Is the new proposed SeqST still applicable to implicit functional relationships in engineering without prior information? In the future work, the performance of the new proposed SeqST under various types of optimal experimental designs, such as uniform designs, orthogonal arrays, D-optimal designs, and various types of machine learning modeling techniques, will be investigated.

- Elsawah 2022d (cf. its Sect. 5) presented a mixture factor-weight WD (MFWWD) as a new criterion for constructing new uniform mixture factor-weight experimental designs (training and testing datasets) when the input factors are not equally important. A comparison study between the classical SinST using the new uniform mixture factor-weight experimental designs and the new proposed SeqST using classical uniform designs in all of its stages will be investigated in the future work.

Acknowledgements The authors greatly appreciate valuable comments and suggestions of the Editor and referees that significantly improved the paper. Elsawah greatly appreciates the kind support of Prof. Kai-Tai Fang.

Funding Elsawah's work was supported by the UIC Research Grants with number of (R201810, R201912 and R202010); the Curriculum Development and Teaching Enhancement with number of (UICR0400046-21CTL); the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College with number of (2022B1212010006); and the Guangdong Higher Education Upgrading Plan (2021-2025) with number of (UIC R0400001-22).

Data availability All data generated or analyzed during this study are included in this article.

Declarations

Conflict of interest There is no conflict of interest.

References

Andre J, Siarry P, Dognon T (2000) An improvement of the standard genetic algorithm fighting premature convergence. Adv Eng Softw 32(1):49–60

Balestrassi PP, Popova E, Paiva AD, Lima JM (2009) Design of experiments on neural network's training for nonlinear time series forecasting. Neurocomputing 72(46):1160–1178

Diwekar UM, Kalagnanam JR (1997) Efficient sampling technique for optimization under uncertainty. AIChE J 43(2):440–447

Elsawah AM (2017a) A closer look at de-aliasing effects using an efficient foldover technique. Statistics 51(3):532–557

Elsawah AM (2017b) A powerful and efficient algorithm for breaking the links between aliased effects in asymmetric designs. Aust N Z J Stat 59(1):17–41

Elsawah AM (2021a) Multiple doubling: a simple effective construction technique for optimal two-level experimental designs. Stat Pap 62:2923–2967

Elsawah AM (2021b) An appealing technique for designing optimal large experiments with three-level factors. J Comput Appl Math 384:113164

Elsawah AM (2022a) A novel non-heuristic search technique for constructing uniform designs with a mixture of two- and four-level factors: a simple industrial applicable approach. J Korean Stat Soc 51:716–757

Elsawah AM (2022b) Designing optimal large four-level experiments: a new technique without recourse to optimization softwares. Commun Math Stat 10:623–652

Elsawah AM (2022c) Improving the space-filling behavior of multiple triple designs. Comput Appl Math 41:180

Elsawah AM (2022d) Novel techniques for performing follow-up experiments based on prior information from initial-stage experiments. Statistics 56(5):1133–1165

Elsawah AM (2023a) A novel hybrid algorithm for designing mixed three- and nine-level experiments without modeling assumptions. Commun Stat Simul Comput. https://doi.org/10.1080/03610918.2023.2269323

Elsawah AM (2023b) A novel doubling-tripling-threshold accepting hybrid algorithm for constructing asymmetric space-filling designs. J Korean Stat Soc. https://doi.org/10.1007/s42952-023-00232-5

Elsawah AM, Gong Y (2023) A new non-iterative deterministic algorithm for constructing asymptotically orthogonal maximin distance Latin hypercube designs. J Korean Stat Soc 52:621–646

Elsawah AM, Vishwakarma GK (2022) A systematic construction approach for nonregular fractional factorial four-level designs via quaternary linear codes. Comput Appl Math 41:323

Fang KT, Li RZ, Sudjianto A (2006) Design and modeling for computer experiments. Chapman and Hall/CRC, New York

Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh

Goksungur YS, Dagbagli AU, Guvenc U (2005) Optimisation of pullulan production from synthetic medium by aureobsidium pullulans in a stirred tank reactor by response surface methodology. J Chem Technol Biotechnol 80(7):819–827

Hammersley JM (1960) Monte Carlo methods for solving multivariate problems. Ann N Y Acad Sci 86(3):844–874

Husslage BGM, Rennen G, van Dam ER, Hertog D (2011) Space-filling Latin hypercube designs for computer experiments. Optim Eng 12(4):611–630

Iman RL, Conover WJ (1980) Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. Commun Stat Theory Methods 17:1749–1842

Iordanis I, Koukouvinos C, Silou I (2022) Classification accuracy improvement using conditioned Latin hypercube sampling in supervised machine learning. In: 12th international conference on dependable systems, services and technologies (DESSERT), Athens, Greece

Lujan-Moreno GA, Howard PR, Rojas OG, Montgomery DC (2018) Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. Expert Syst Appl 109:195–205

Mark L (2009) The mathematical mechanic: using physical reasoning to solve problems. Princeton University Press, Princeton

Mckay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(2):239–245

Mohamed HS, Elsawah AM, Shao YB, Wu CS, Bakri M (2023) Analysis on the shear failure of HSS S690-CWGs via mathematical modelling. Eng Fail Anal 143:106881

Ortiz-Rodriguez JM, Martinez-Blanco MR, Vega-Carrillo HR (2006) Robust design of artificial neural networks applying the Taguchi methodology and DoE. In: Electronics, robotics and automotive mechanics conference (CERMA'06), pp 131–136. https://doi.org/10.1109/CERMA.2006.83

Owen AB (1992) Orthogonal arrays for computer experiments, integration and visualization. Stat Sin 2:439–452

Packianather MS, Drake PR, Rowlands H (2000) Optimizing the parameters of multilayered feedforward neural networks through Taguchi design of experiments. Qual Reliab Eng Int 16(6):461–473

Prasath BB, Elsawah AM, Liyuan Z, Poon K (2021) Modeling and optimization of the effect of abiotic stressors on the productivity of the biomass, chlorophyll and lutein in microalgae *Chlorella pyrenoidosa*. J Agric Food Res 5:100163

Prasath BB, Zahir M, Elsawah AM, Raza M, Lecong C, Chutian S, Poon K (2022) Statistical approaches in modeling of the interaction between bacteria and diatom under a dual-species co-cultivation system. J King Saud Univ Sci 34(1):101743

Ranjan R, Bingham D, Michailidis G (2008) Sequential experiment design for contour estimation from complex computer codes. Technometrics 50:527–541

Salmaso L, Pegoraro L, Giancristofaro RA, Ceccato R, Bianchi A, Restello S, Scarabottolo D (2022) Design of experiments and machine learning to improve robustness of predictive maintenance with application to a real case study. Commun Stat Simul Comput 51(2):570–852

Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3(3):210–229

Sikirica A, Grbcic L, Kranjcevic L (2023) Machine learning based surrogate models for microchannel heat sink optimization. Appl Therm Eng 222:119917

Staelin C (2003) Parameter selection for support vector machines. Hewlett-Packard Company, London

Sukthomya W, Tannock J (2005) The optimisation of neural network parameters using Taguchi's design of experiments approach: an application in manufacturing process modelling. Neural Comput Appl 14(4):337–344

Viana FAC (2013) Things you wanted to know about the Latin hypercube design and were afraid to ask. In: 10th world congress on structural and multidisciplinary optimization

Zhang S, Feng G, Yuan F, Guo S (2022) Twin support vector regression model based on heteroscedastic Gaussian noise and its application. IEEE Access 10:111738–111748

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.