Journal of Statistical Planning and Inference 164 (2015) 10–26

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

The depth-design: An efficient generation of high dimensional computer experiments

Martin Piffl[a,*], Ernst Stadlober[b]

[a] AVL List GmbH, Hans-List-Platz 1, 8020 Graz, Austria

[b] Graz University of Technology, Institute of Statistics, NAWI Graz, Kopernikusgasse 24/III, 8010 Graz, Austria

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form 5 March 2015
Accepted 6 March 2015
Available online 16 March 2015

Keywords:
Multiple linear regression
Screening designs
Total sensitivity index
Space filling designs
Median-oriented quantiles
Depth functions

ABSTRACT

This paper provides an approach on how to generate representative experiments for the investigation of a model based system or process, depending on quantitative variables, when the number of experiments N is limited ($25 \leq N \leq 500$). An exemplified overview of known screening designs that are suitable for quadratic response surfaces possibly depending on $k \geq 50$ factors is given. The relevance of these factors is measured by a sensitivity index, which is based on corresponding sums of squares of the underlying linear, quadratic as well as the linear two way interaction effects. Bearing in mind the sparsity-of-effects principle, we expect the process or system to be dominated only by a minority of the factors ($k_r \leq 10$) assumed. Among other space filling designs we especially investigate the very efficient Latin Hypercube Design in terms of its capability to represent a multidimensional distribution with its experiments. We use the theory of median-oriented quantiles and depth functions to assess this capability and to introduce our new space filling design approach, the Depth-Design. On the example of the multivariate normal distribution we demonstrate that our Depth-Design represents a multidimensional distribution with much less experiments in comparison to the Latin Hypercube design.

1. Introduction

Computers have become an indispensable tool in all technical fields where complex processes need to be investigated and optimized. Although simulation models are becoming more and more accurate, the associated increase in complexity of these models has hesitated much faster computation times and has caused confusing frameworks over the last twenty years (cf. Currin et al., 1991; Siebertz et al., 2010).

Design of experiments (DOE) minimizes the required effort of simulations (or "simply" experiments) to be run for an investigation of a response variable Y within given system boundaries that enclose the so called feature space. DOE additionally generates the experiments necessary to optionally set up a fast predicting regression model of Y, where complex system interactions can be easily tracked. These advantages make DOE a very attractive tool, which can be used to overcome the difficulties coming along with present day computer simulation models (Montgomery, 2012). Nevertheless, computer simulation results are frequently non-linearly determined by a large number ($k \geq 50$) of factors $X_1, \ldots, X_k$ so that most usual DOE approaches, like fractional factorial designs or the central composite designs, would exhibit a too large number of experiments, or suffer from ambiguous possibilities of interpretation. Given that the feature space is subject

* Corresponding author. Tel.: +43 660 657 37 38.
E-mail address: martin@piffl.com (M. Piffl).

http://dx.doi.org/10.1016/j.jspi.2015.03.002
0378-3758/

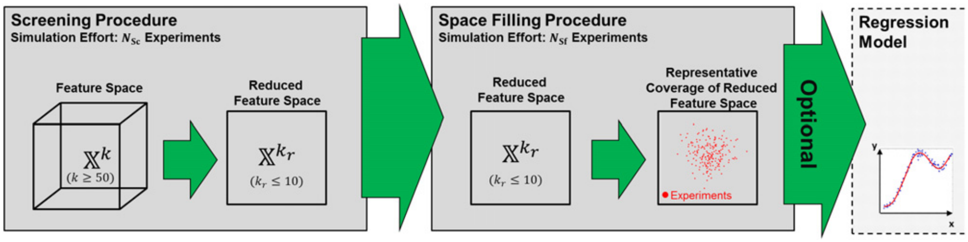

Fig. 1. Guideline to generate representative experiments.

to a multidimensional distribution, a representative experimental coverage is a challenge to be additionally mastered. For these reasons, this paper provides a guideline on how to identify the most significant factors $X_1, \ldots, X_{k_r}$ for Y, and how to represent the distribution of the remaining feature space with a limited number of experiments.

In Section 2 we compare 3-level screening designs and emphasize the power of the Definitive Screening Design of Jones and Nachtsheim (2011) in the context of computer simulation models. We recall the total sensitivity index of Homma and Saltelli (1996) in order to identify the most significant factors. Thereafter, in Section 3 we compare the capability of the Latin Hypercube design to generate representative experiments for a multidimensional distribution with the aid of median-oriented quantiles and depth functions, discussed in detail by Serfling (2010). With regard to the results obtained we develop our new space filling design approach, the Depth-Design, which is presented in Section 4. In the end we conclude our findings and present our recommendations.

2. Generation of representative experiments

In this paper we assume that the random vector $\mathbf{X} = (X_1, \ldots, X_k) \sim F^k$ represents all feasible combinations of input factors of a computer simulation model, whereas $X_1, \ldots, X_k$ are considered to be quantitative variables. The domain of the multidimensional distribution F^k is denoted as the feature space $\mathbb{X}^k \subseteq \mathbb{R}^k$, which contains all feasible experiments realized by vectors $\mathbf{x} = (x_1, \ldots, x_k)$ of $\mathbf{X}$. Furthermore, we are interested in the simulation results of one response variable Y, obtained by experiments $\mathbf{x} \in \mathbb{X}^k$.

Given that only one response variable Y is of interest, and the number of experiments is constrained (i.e. $25 \leq N \leq 500$), we propose to generate representative experiments not only with regard to the feature space distribution F^k, but also with regard to the response Y. This is carried out as process (see Fig. 1) consisting of a screening and a space filling procedure. We use the rule of thumb that most processes or systems are dominated by a few factors. A preliminary screening procedure, using a portion N_{Sc} of the number feasible experiments N, can cheaply reduce all considered factors $X_1, \ldots, X_k$ to the most significant factors $X_1, \ldots, X_{k_r}$ (usually $k_r \leq 10$) for Y. As a consequence, it is possible to reduce the feature space $\mathbb{X}^k$ to a subspace $\mathbb{X}^{k_r}$. The associated reduction in dimension of the feature space eases the experimental coverage of the resulting multidimensional distribution F^{k_r} with N_{Sf} experiments during the subsequent space filling procedure. Once, representative data is gathered, there exists the option to build a regression model. Still, it remains to be clarified, which type of screening design and which type of space filling design are most appropriate for F^{k_r}, when the number of experiments $N = N_{Sc} + N_{Sf}$ is limited.

2.1. Screening designs

Screening designs enable the identification of the most relevant factors $X_1, \ldots, X_k$ in terms of the response Y with comparable low simulation effort. Given a fixed number of factors k to investigate, screening designs mainly differ in the number of experiments N_{Sc} and in the interpretability of the estimated effects. While 2-level screening designs, share the idea of comparing the results of two different levels " -1" and "$+1$" (extreme case scenarios), 3-level screening designs additionally consider the factors at a center level "0" in order to detect possible curvature in the relationship between $X_1, \ldots, X_k$ and Y.

For the screening procedure factors $X_1, \ldots, X_k$ are assumed as independent random variables. This simplification may be especially feasible for computer simulation models, and may only lead to a larger number of selected factors k_r obtained by the screening procedure due to neglected correlation structure. If the independent setting of two or more factors is not possible, it is suggested to successively neglect such factors until an independent consideration becomes possible. It is proposed to standardize the feasible ranges of $X_1, \ldots, X_k$, which need to be chosen after good engineering judgment, to $[-1, +1]$ so that eventually $\mathbb{X}^k = [-1, +1]^k$. Scientists and engineers do often feel more comfortable with 3-level screening designs, because they tend to expect a substantial non-linear relationship between the factors $X_1, \ldots, X_k$ and the response Y. As a result, the straightforward application of 3-level full factorial designs (3^k designs) is not possible, because the number of experiments N_{Sc} explodes when $k \geq 50$. Therefore, we shortly discuss the following alternative 3-level screening designs.

1. 3^{k-p} fractional factorial design (Montgomery, 2012)
2. Box–Behnken Design (Box and Behnken, 1960)