Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

CONTROLLED MORRIS METHOD: A NEW DISTRIBUTION-FREE SEQUENTIAL TESTING PROCEDURE FOR FACTOR SCREENING

Wen Shi

School of Logistics and Engineering Management
Hubei University of Economics
Wuhan 430205, CHINA

Xi Chen

Department of Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061, USA

ABSTRACT

Morris's elementary effects method (MM) has been known as a model-free factor screening approach especially well-suited when the number of factors is large or when the computer model is computationally expensive to run. In this paper, we propose the controlled Morris method (CMM) that acts in a sequential manner to keep the computational effort down to a minimum. The sequential probability ratio test-based multiple testing procedure adopted by CMM enables to identify the factors with significant main and/or interaction effects while controlling Type I and Type II familywise error rates at desired levels. A numerical example is provided to demonstrate the efficacy and efficiency of CMM.

1 INTRODUCTION

Factor screening refers to the process of identifying, through design of experiments, statistical modeling and sampling, those factors that have a significant influence on the model output. Proposed for factor screening in the context of deterministic computer experiments by Morris (1991), Morris's elementary effects method (MM) has been known as a *model-free* approach particularly well-suited when the number of factors is relatively large or when the computer model is computationally expensive to run. Recently, Campolongo and Braddock (1999) and Cropp and Braddock (2002) extend standard MM by providing estimates of two-factor interaction effects. Campolongo et al. (2007) propose to use normalized elementary effects as compared to those used in standard MM so that the performance of MM can be more robust. Boukouvalas et al. (2014) propose to implement MM in a sequential way so that factors having nonlinear effects can be identified more efficiently. Most recently, Fédou and Rendas (2015) present a fast mixed effects screening method that enables efficient estimation of the interaction graph of factors. Shi et al. (2016) propose an effective error control mechanism for controlling the overall false discovery rate achieved by MM, and reveal its connections with other screening methods such as sequential bifurcation (e.g., Bettonvil and Kleijnen 1997, Wan et al. 2010).

Despite the aforementioned improvements made to standard MM, little attention has been given to establishing an adaptive sampling procedure for MM with a rigorous statistical guarantee on its screening performance. In this paper, we propose the controlled Morris method (**CMM**) that adopts a novel distribution-free sequential probability ratio test (SPRT)-based multiple testing procedure for identifying factors that have significant main and/or interaction effects while ensuring the Type I and II familywise error rates controlled at desired levels.

While SPRT-based procedures have been proposed for multiple hypothesis testing, most of them rely on the assumption that the underlying distribution from which the observations are sampled is known (Wald 1992; De and Baron 2012b; De and Baron 2012a; Bartroff and Song 2014); among those distributions stipulated, Gaussian is particularly popular (Wan et al. 2010; Ankenman et al. 2014; Shi et al. 2014).

However, the distribution is typically unknown in practice, or even if known it cannot be specified by a simple distribution function. Despite the significant role played by nonparametric estimation methods in modern statistics, nonparametric SPRT-based hypothesis testing procedures have been rarely studied in the literature, to the best of our knowledge. Antoniak and Dillard (1968) and Yu and Su (2004) are among the few that have relaxed the distribution-known assumption, and they investigate Wilcoxon signed rank statistics but still have to assume that the underlying distribution is symmetric around the median. The SPRT-based sequential multiple testing procedure adopted by CMM, on the other hand, is fully distribution free, thanks to the use of online kernel density estimation.

The remainder of this paper is organized as follows. In Section 2, we give a brief review of the Morris's elementary effects method. In Section 3, we provide details on the controlled Morris method (CMM). Section 4 provides a numerical evaluation of CMM. Section 5 concludes the paper.

2 A REVIEW ON MORRIS METHOD

The Morris's elementary effects method (MM) is originally proposed for factor screening in the context of deterministic computer experiments (Morris 1991). Suppose that there are k factors in total in the simulation model and each factor is scaled to take values from $[0,1]$. For the purpose of factor screening, MM considers varying the value of each factor across p pre-selected levels in $[0,1]$; that is, the experimental region Ω for MM is a k-dimensional p-level grid in $[0,1]^k$.

Let $\mathsf{Y}(\mathbf{x})$ be the output obtained by running a deterministic computer experiment at factor combination $\mathbf{x} = (x_1,\ldots,x_k)^\top \in \Omega$. The elementary effect of the jth factor at $\mathbf{x}$ is defined as

$$d_j(\mathbf{x}) = \frac{\mathsf{Y}(\mathbf{x}+e_j\Delta)-\mathsf{Y}(\mathbf{x})}{\Delta}, \quad j=1,\ldots,k, \tag{1}$$

where e_j denotes the unit vector in the direction of the jth axis; Δ is a predefined integer multiple of $1/(p-1)$ such that $\mathbf{x}+e_j\Delta \in \Omega$. Hence, the jth factor x_j assumes values in $\{0,1/(p-1),2/(p-1),\ldots,1\}$. Intuitively speaking, $d_j(\mathbf{x})$ can be thought of as the partial derivative of $\mathsf{Y}(\mathbf{x})$ with respect to x_j when Δ is small (Woods and Lewis 2016).

The elementary effects corresponding to the jth factor, $d_j(\mathbf{x})$, follow a finite distribution, F_j, which can be obtained by randomly sampling the factor combination $\mathbf{x}$ from Ω. The number of elements of F_j is then $p^{k-1}[p-\Delta(p-1)]$, where p^{k-1} is the number of factor combinations formed by the remaining $k-1$ factors, and $p-\Delta(p-1)$ is the number of possible levels that factor j can take to obtain elementary effects; for example, when $\Delta = 1/(p-1)$, factor j can only have $p-1$ levels (i.e., the level corresponding to $x_j = 1$ is excluded). A recommended choice of p is even and $\Delta = p(2(p-1))^{-1}$ (Morris 1991).

A highly centralized distribution F_j suggests a consistent importance of factor j across the experimental region Ω, and a highly decentralized distribution indicates a strong dependence of factor j on the other factors (i.e., nonlinear or interaction effects may be present). MM determines the importance of the jth factor in terms of two measures, μ_j and σ_j, respectively, the mean and standard deviation of F_j. To estimate these two measures, MM samples N elementary effects from F_j via a sampling design that generates N trajectories in Ω (Morris 1991), and the ith trajectory provides k elementary effects estimates $d_{j;i}$ for $j=1,2,\ldots,k$. The following unbiased estimators of the mean and variance of F_j are then used to assess the importance of factor j,

$$\widehat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} d_{j;i}, \tag{2}$$

$$\widehat{\sigma}_j^2 = \frac{1}{N-1}\sum_{i=1}^{N} (d_{j;i}-\widehat{\mu}_j)^2 \tag{3}$$

where recall that $d_{j;i}$ denotes the ith elementary effect randomly generated for factor j from F_j.

In practice, Morris (1991) recommends to use a graph plotting $\widehat{\mu}_j$ vs. $\widehat{\sigma}_j$ with two lines corresponding to $\widehat{\mu}_j = \pm 2\widehat{\sigma}_j/\sqrt{N}$ for assessing the importance of factor j: If the point $(\widehat{\mu}_j, \widehat{\sigma}_j)$ locates outside the wedge formed by the two lines, then factor j is deemed important. Such a practice, however, is more of a commonsense rule than a rigorously justified screening method.

To generate elementary effects from F_j, Morris introduces using *sampling matrix* and *design matrix*. For example,

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{4}$$

is a $(k+1) \times k$ sampling matrix, which consists of a $1 \times k$ vector of zeros in the first row and a $k \times k$ lower triangular matrix whose entries below the main diagonal are all ones. The design matrix corresponding to $\mathbf{B}$ is $\Delta\mathbf{B} := \Delta \times \mathbf{B}$. Though a sampling matrix is easier to understand, its corresponding design matrix is the one actually used for running simulation experiments.

Notice that $\mathbf{B}$ (or equivalently, $\Delta\mathbf{B}$) can only generate one elementary effect for each factor, which holds true for any *random form* (also called *random orientation* or *trajectory*) of $\Delta\mathbf{B}$, generically denoted by $\mathbf{B}^+$. For example, when $k = 3$ and $p = 4$ (so that $\Delta = p\left(2\left(p-1\right)\right)^{-1} = 2/3$), one random form of $\Delta\mathbf{B}$ can be

$$\mathbf{B}^+ = \begin{bmatrix} 2/3 & 0 & 1 \\ 2/3 & 2/3 & 1 \\ 0 & 2/3 & 1 \\ 0 & 2/3 & 1/3 \end{bmatrix}.$$

Morris (1991) provides a special algorithm to convert a sampling matrix $\mathbf{B}$ to a random design matrix $\mathbf{B}^+$ given specified k and p. To obtain N ($N \geq 2$) *independent* elementary effects for each factor, Morris suggests to use N random forms of $\Delta\mathbf{B}$ and form the ultimate design matrix for running simulation experiments as $\left(\mathbf{B}_1^+, ..., \mathbf{B}_N^+\right)^\top$.

3 THE CONTROLLED MORRIS METHOD

In this section we describe the controlled Morris method (CMM) that acts in a sequential manner to keep the number of simulation runs down to a minimum, while identifying the factors that have significant main and/or interaction effects with Type I and Type II familywise error rates controlled at desired levels.

3.1 The Set-Up of CMM

Suppose that there are k factors in a deterministic simulation model. Let $\mathbf{d}_j = \{d_{j;1}, d_{j;2}, \ldots\}$ denote the set of elementary effects generated sequentially for factor j, where $d_{j;1}, d_{j;2}, \ldots$ are independent and identically distributed (i.i.d.) following a common distribution given by F_j whose corresponding probability density function (PDF) is given by f_j, for $j = 1, 2, \ldots, k$. The goal is to determine whether each factor has a significant main or interaction effect or both, via simultaneously performing the following $2k$ individual hypothesis tests.

For a given factor j, we want to determine if it has a significant main effect using the following one-sided hypothesis test:

$$H_l : |\mu_j| \leq \Delta_{\mathrm{EE}}^{(0)} \quad \text{vs.} \quad G_l : |\mu_j| \geq \Delta_{\mathrm{EE}}^{(1)}, \quad \text{for } l = 2j-1; j = 1, 2, \ldots, k, \tag{5}$$

where the null and alternative are respectively denoted by H_l and G_l; $\Delta_{\text{EE}}^{(0)}$ and $\Delta_{\text{EE}}^{(1)}$ are respectively the user specified parameters that define the thresholds of nonsignificant and significant main effects and they satisfy $0 \le \Delta_{\text{EE}}^{(0)} \le \Delta_{\text{EE}}^{(1)}$.

We note that (5) is equivalent to the following two alternative hypotheses tests depending on the sign of μ_j:

$$\begin{cases} H_l : 0 \le \mu_j \le \Delta_{\text{EE}}^{(0)} \ \text{ vs. } \ G_l : \mu_j \ge \Delta_{\text{EE}}^{(1)}, \text{ if } \mu_j \ge 0; \\ H_l : -\Delta_{\text{EE}}^{(0)} \le \mu_j \le 0 \ \text{ vs. } \ G_l : \mu_j \le -\Delta_{\text{EE}}^{(1)}, \text{ if } \mu_j < 0. \end{cases} \tag{6}$$

That is, when $\mu_j \ge 0$, we consider the main effect of factor j to fall into one of the following two categories: (i) nonsignificant, if $\mu_j \le \Delta_{\text{EE}}^{(0)}$; (ii) significant, if $\mu_j \ge \Delta_{\text{EE}}^{(1)}$; moreover, if $\mu_j \in (\Delta_{\text{EE}}^{(0)}, \Delta_{\text{EE}}^{(1)})$, then we want to provide a sound power to identify it as significant. Analogous description holds for the alternative case where $\mu_j < 0$. Notice that the sign of μ_j can be observed through that of $\hat{\mu}_j$ calculated with a sample of elementary effects generated for factor j (Campolongo et al. 2007), and it may change as new elementary effects continue to arrive.

Similarly, we use the following one-sided hypothesis test to determine if factor j has a significant interaction effect with other factors:

$$H_l : \sigma_j \le \Delta_{\text{IE}}^{(0)} \ \text{ vs. } \ G_l : \sigma_j \ge \Delta_{\text{IE}}^{(1)}, \quad \text{for } l = 2j; j = 1, 2, \ldots, k, \tag{7}$$

where $\Delta_{\text{IE}}^{(0)}$ and $\Delta_{\text{IE}}^{(1)}$ respectively denote the thresholds of nonsignificant and significant interaction effects that satisfy $0 \le \Delta_{\text{IE}}^{(0)} \le \Delta_{\text{IE}}^{(1)}$. The interaction effect of factor j will be classified into the following two categories: (i) nonsignificant, if $\sigma_j \le \Delta_{\text{IE}}^{(0)}$; and (ii) significant, if $\sigma_j \ge \Delta_{\text{IE}}^{(1)}$; moreover, if $\sigma_j \in (\Delta_{\text{IE}}^{(0)}, \Delta_{\text{IE}}^{(1)})$, then we want to provide a sound power to identify it as significant.

For notational convenience, let $\theta \triangleq (\theta_1, \theta_2, \ldots, \theta_{2k})^\top$ be the vector $(\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, \mu_k, \sigma_k)^\top$ whose dimensions are $2k \times 1$. That is, for $l \in \{1, 2, \ldots, 2k\}$, θ_l denotes the main effect (respectively, interaction effect) for factor $\lceil l/2 \rceil$ if l is odd (resp., even). Denote $\mathscr{H}(\theta) = \{l \in \{1, 2, \ldots, 2k\} : \theta_l \in H_l\}$ the set of indices whose corresponding null hypotheses are true; that is, factor $\lceil l/2 \rceil$ has a truly nonsignificant main or interaction effect. Let $\mathscr{G}(\theta) = \{l \in \{1, 2, \ldots, 2k\} : \theta_l \in G_l\}$ be the set of indices whose corresponding null hypotheses are false; that is, factor $\lceil l/2 \rceil$ has a truly significant main or interaction effect.

The aforementioned factor screening problem naturally falls into the multiple hypothesis test setting, where it is vital to control the error rates achieved especially when simultaneously testing a considerably large number of hypotheses (i.e., when k is large). Upon a stopping rule T is given which will be specified later, with decisions regarding acceptance or rejection of each of the $2k$ null hypotheses, CMM aims to control the *Type I and Type II familywise error rates* (De and Baron 2012b, Bartroff and Song 2014) by guaranteeing that

$$\begin{aligned} \text{FWE}_{\text{I}}(\theta) &= \mathsf{P}\{H_l \text{ is rejected for some } l \in \mathscr{H}(\theta)\} \le \alpha \\ \text{FWE}_{\text{II}}(\theta) &= \mathsf{P}\{G_l \text{ is rejected for some } l \in \mathscr{G}(\theta)\} \le \beta, \end{aligned} \tag{8}$$

where $\alpha, \beta \in (0, 1)$ are two user-specified parameters, in addition to $\Delta_{\text{EE}}^{(0)}$, $\Delta_{\text{EE}}^{(1)}$, $\Delta_{\text{IE}}^{(0)}$, and $\Delta_{\text{IE}}^{(1)}$. Notice that the quantity $1 - \text{FWE}_{\text{II}}(\theta)$ is also known as "*familywise power*"; equivalently, CMM aims to provide $\gamma \triangleq 1 - \beta$ familywise power for the entire factor screening procedure.

3.2 Description of the CMM's Sequential Multiple Testing Procedure

To effectively control the Type I and Type II familywise error rates simultaneously, CMM adopts a novel distribution-free sequential probability ratio test (SPRT)-based multiple testing procedure. Essentially, this procedure can be thought of as an ensemble of $2k$ individual SPRTs for identifying the significance of the main and interaction effects associated with each of the k factors.

We now describe the procedure in terms of stages of sampling, between which accept/reject decisions are made to each hypothesis test. Without loss of generality, let n denote the cumulative sample size of

elementary effects collected for any active test (i.e., the H_l for which no decision has been reached yet) that have up to and including the current stage. For the lth ($l = 1,2,\ldots,2k$) pair of hypotheses given by either (6) or (7), the testing of H_l vs. G_l is based on the following test statistic (De and Baron 2012a, De and Baron 2012b, Bartroff and Song 2014, Wang and Wan 2014),

$$\Lambda_l(n) = \sum_{i=1}^{n} \big(\log f_l(d_{j,i}|\theta_l \in G_l) - \log f_l(d_{j,i}|\theta_l \in H_l)\big), \qquad \text{with } j = \lceil l/2 \rceil, \tag{9}$$

where $f_l(\cdot|\theta_l \in H_l)$ and $f_l(\cdot|\theta_l \in G_l)$ are, respectively, the PDFs of elementary effects for factor $\lceil l/2 \rceil$ given that H_l and G_l are true, for $\ell = 1,2,\ldots,2k$. Notice that $\Lambda_l(n)$ is known as *Kullbak-Leibler information numbers* (De and Baron 2012b), and it is used for measuring the "distance" between two probability measures defined on a common measurable space (Dykstra 2005).

Let α_l and β_l be the prescribed levels of the Type I and Type II error rates to achieve for the lth test according to Wald's SPRT, and let a_l and b_l be the upper and lower stopping boundaries, respectively. Wald's SPRT for the single lth hypothesis H_l rejects it (i.e., chooses G_l) upon obtaining n elementary effects if $\Lambda_l(n) \geq a_l$, accepts it (i.e., chooses H_l) if $\Lambda_l(n) \leq b_l$, and continues sampling elementary effects for factor $\lceil l/2 \rceil$ if $\Lambda_l(n) \in (a_l, b_l)$.

The sequential multiple testing procedure of CMM starts with an initial sample size of n_0 elementary effects, and examines whether the resulting test statistic $\Lambda_l(n_0)$ crosses one of the two stopping boundaries (i.e., a_l or b_l), which leads to the decision of declaring effect l significant (respectively, the upper boundary is crossed) or nonsignificant (resp., the lower boundary is crossed). If neither one is crossed, then the procedure continues sampling one elementary effect for factor $\lceil l/2 \rceil$ per stage until $\Lambda_l(n)$ escapes from the continue-sampling region by crossing one of the two boundaries (i.e., $\Lambda_l(n) \notin (b_l, a_l)$). Let T_l be the stopping time of the lth hypothesis test, namely,

$$T_l = \inf\{n \geq n_0 : \Lambda_l(n) \notin (b_l, a_l)\}. \tag{10}$$

It is clear that the procedure continues sampling until all $2k$ tests reach decisions, and the stopping time of the entire procedure follows as

$$T = \inf\left\{ n \geq n_0 : \bigcap_{l=1}^{2k} \{\Lambda_l(n) \notin (b_l, a_l)\} \right\}. \tag{11}$$

Lemma 2 of De and Baron (2012b) suggests that the lth test can control the Type I and Type II error probabilities by using the upper and lower stopping boundaries given by a_l and b_l because

$$\begin{aligned} &\mathsf{P}\{H_l \text{ is rejected for some } l \in \mathscr{H}(\theta)\} \leq \mathsf{P}\{\Lambda_l(n) \geq a_l \mid l \in \mathscr{H}(\theta)\} \leq e^{-a_l}, \\ &\mathsf{P}\{H_l \text{ is accepted for some } l \in \mathscr{G}(\theta)\} \leq \mathsf{P}\{\Lambda_l(n) \leq b_l \mid l \in \mathscr{G}(\theta)\} \leq e^{b_l}. \end{aligned} \tag{12}$$

Therefore, by setting the two boundaries respectively as $a_l = -\log \alpha_l$ and $b_l = \log \beta_l$, we can control the Type I and Type II error probabilities conservatively at levels $\alpha_l = e^{-a_l}$ and $\beta_l = e^{b_l}$ for $l = 1,2,\ldots,2k$. Furthermore, it follows from Theorem 1 of De and Baron (2012b) that the sequential multiple testing procedure with the stopping time given by (11) can control $\mathrm{FWE}_{\mathrm{I}}(\theta)$ and $\mathrm{FWE}_{\mathrm{II}}(\theta)$ respectively at levels α and β, if we set $\alpha_l = \alpha/2k$ and $\beta_l = \beta/2k$ thanks to the Bonferroni's inequality.

3.3 Online Kernel Density Estimation

In this section we propose to approximate $f_l(\cdot)$ used in (9) via online kernel density estimation. We then use the estimated density function $\widehat{f_l}(\cdot)$ to derive a distribution-free SPRT test statistic according to (9), denoted by $\widehat{\Lambda_l}(n)$, by replacing $f_l(\cdot)$ with the estimate $\widehat{f_l}(\cdot)$.

As an indispensable nonparametric estimation tool, kernel density estimation (KDE) specializes in capturing the behaviors of the distribution of interest without imposing any parametric assumptions (Härdle 1990; Zhang and Wang 2014; Parpas et al. 2015); more importantly, it is more flexible to accommodate complex settings such as the underlying density is multimodal (Han et al. 2008). Though different KDE methods can be employed under the CMM framework, in this paper we restrict our attention to the *online (sequential) kernel density estimation* (oKDE) method proposed by Kristan et al. (2010) and Kristan et al. (2011). oKDE is a type of machine learning approach, which updates $\widehat{f}_l(\cdot)$ each time upon receiving a new observation so as to continuously provide accurate estimation from the data observed thus far. Notice that this feature suits the setting of sequential sampling for multiple hypothesis testing very well.

For the lth effect, we obtain an empirical probability density function (EPDF), denoted by $\widehat{f}_l$, based on the set of elementary effects for factor $\lceil l/2 \rceil$ via oKDE, for $l = 1,2,\ldots,2k$. For notational simplicity, we drop the subscript l in the following discussion, as this approach can be applied to any of the $2k$ effects. In short, oKDE provides an initial estimate $\widehat{f}$ based on a sample of n_0 elementary effects collected in the pilot stage, and proceeds following three steps listed below upon receiving a new elementary effect in each subsequent sampling stage: (1) update $\widehat{f}$ with the new observation; (2) re-estimate the *optimal bandwidth* used by oKDE; and (3) refine and compress the estimate $\widehat{f}$. The necessity of the last step will be explained shortly.

Specifically, using a sample of n elementary effects for a given factor, say, $\{d_i, i = 1,2,\ldots,n\}$, the EPDF $\widehat{f}$ can be given by an n-component *Gaussian mixture model* as follows,

$$\widehat{f}_n(d) = \sum_{i=1}^{n} \omega_i \mathscr{K}_{h_i}(d - d_i), \tag{13}$$

where ω_i denotes the weight of the ith component, and $\mathscr{K}_{h_i}(d-d_i)$ is a Gaussian kernel, which is defined as

$$\mathscr{K}_{h_i}(d-d_i) = (2\pi h_i^2)^{-\frac{1}{2}} \exp\left(-\frac{(d-d_i)^2}{2h_i^2}\right), \tag{14}$$

and d_i and h_i denote, respectively, the center and *bandwidth* of the ith Gaussian kernel, for $i = 1,2,\ldots,n$.

With the n_0 elementary effects obtained in the pilot stage, the kernel density estimate consisting of n_0 evenly weighted kernels constructed using an equal bandwidth h_{n_0} can be expressed as

$$\widehat{f}_{n_0}(d) = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathscr{K}_{h_{n_0}}(d-d_i) = \frac{1}{n_0}\sum_{i=1}^{n_0} (2\pi h_{n_0}^2)^{-\frac{1}{2}} \exp\left(-\frac{(d-d_i)^2}{2h_{n_0}^2}\right) \tag{15}$$

We note that due to the absence of sufficient information about f in the pilot stage, an equal weight and identical bandwidth are used in all the components of $\widehat{f}_{n_0}(d)$.

Without loss of generality, suppose that we have obtained $n-1$ elementary effects ($n \geq n_0+1$) and derived the EPDF $\widehat{f}_{n-1}(d)$. Upon receiving an additional elementary effect d_n, the EPDF can be updated according to the following expression,

$$\widehat{f}_n(d) = \left(1 - \frac{1}{n-1}\right)\widehat{f}_{n-1}(d) + \frac{1}{n}\mathscr{K}_{h_n}(d-d_n), \tag{16}$$

where the new observation is assigned a weight of n^{-1}. It is obvious that the estimation accuracy achieved by $\widehat{f}_n(d)$ depends heavily on the choice of bandwidth h_n. By minimizing the asymptotic mean integrated squared error (Kristan et al. 2010), the *optimal bandwidth* can be obtained as

$$h_n^* = \left(\frac{1}{2n\sqrt{\pi}\int f''(x)^2\,dx}\right)^{\frac{1}{5}}, \tag{17}$$

where $f''(\cdot)$ denotes the second derivative of $f(\cdot)$. As n becomes large, $f(d)$ can be approximated well by $\widehat{f}_{n-1}(d)$. Therefore, an estimator $\widehat{h}_n^*$ of h_n^* can be obtained by substituting $\widehat{f}_{n-1}(d)$ into (17); and $\widehat{f}_n(d)$ can be obtained from (16) by using $\widehat{h}_n^*$ in place of h_n^*.

We observe from (16) that as the number of components in $\widehat{f}_n(d)$ increases linearly with the sample size of elementary effects n, the computational complexity approximately scales as $O((n^2-n)/2+n)$, which is mainly attributed to the calculation of bandwidth given by (17) (Kristan et al. 2011). To alleviate this computational burden, a handful of methods have been proposed to reduce (or compress) the number of components in the estimator given (López-Rubio and de Lazcano-Lobato 2008; Deng et al. 2008; Kristan et al. 2010; Kristan et al. 2011). In this paper, we adopt the approach proposed by Kristan et al. (2011) to maintain a comparatively stationary model scale (i.e., use a compressed model). The underlying idea of compression is to identify M ($\leq n$) clusters of similar components in the original kernel estimator, such that components belonging to the mth ($m = 1,2,\ldots,M$) cluster can be well characterized by a single component centered at $\check{d}_m$ in the compressed estimator. Specifically, this approach can help reduce the original n-component model given in (16) to the following one comprised of only M components:

$$\check{f}(d) = \sum_{m=1}^{M} \check{\omega}_m \mathscr{K}_{\check{h}_m}(d - \check{d}_m), \tag{18}$$

where the compressed parameters $\check{\omega}_m = \sum_{i\in\pi(m)} \omega_i$, $\check{d}_m = \check{\omega}_m^{-1} \sum_{i\in\pi(m)} \omega_i d_i$ and $\check{h}_m = \check{\omega}_m^{-1} \sum_{i\in\pi(m)} \omega_i (h_i + d_i^2) - \check{d}_m^2$, and $\pi(m)$ denotes the collection of mth disjoint (compressed) set of indices, for $m = 1,2,\ldots,M$. The key of implementing the compression is to seek an appropriate clustering allocation $\{\pi(m)\}_{m=1}^{M}$, together with the minimum number of clustering number M to use. For the sake of brevity, we omit the details of the compression procedure and refer the interested reader to Algorithm 1 of Kristan et al. (2011).

3.4 Construction of Nonparametric Test Statistics

The task of obtaining a nonparametric estimator of $\Lambda_l(n)$ (given in (9)) to test the lth pair of hypotheses boils down to providing two separate EPDFs *under* the null and alternative hypotheses, namely, $\widehat{f}_l(\cdot|\theta_l \in H_l)$ and $\widehat{f}_l(\cdot|\theta_l \in G_l)$. We note that an estimator $\widehat{f}_l$ directly obtained based on the original sample of elementary effects for factor j, $\mathbf{d}_j = \{d_{j;1}, d_{j;2}, \ldots\}$ with $j = \lceil l/2 \rceil$, may be an appropriate estimator for neither $f_l(\cdot|\theta_l \in H_l)$ nor $f_l(\cdot|\theta_l \in G_l)$.

Inspired by Shi et al. (2016), we transform the original sample of elementary effects $\mathbf{d}_j$ to a sample that complies with a parameter setting that is consistent with a given hypothesis of the lth test. For testing either the main or interaction effect, we apply two transformations to $\mathbf{d}_j$, such that the transformed samples, $\tilde{\mathbf{d}}_l^{(s)}$ for $s = 0$ and 1, complies with the H_l (respectively, G_l) when $s = 0$ (resp., $s = 1$). We note that the superscripts (0) and (1) correspond to the null and alternative hypotheses, respectively.

When testing the significance of the main effect of factor j via the lth hypothesis test with $l = 2j-1$ for $j = 1,2,\ldots,k$, the following transformation equation is applied to $\mathbf{d}_j$ to obtain $\tilde{\mathbf{d}}_l^{(s)}$ for $s = 0$ and 1:

$$\tilde{d}_{l;i}^{(s)} = d_{j;i} - \widehat{\mu}_j(n) + \Delta_{\mathrm{EE}}^{(s)}, \quad i = 1,2,\ldots,n, \tag{19}$$

where n denotes the sample size of $\mathbf{d}_j$ and $\widehat{\mu}_j(n)$ denotes the sample mean of the n original elementary effects obtained for factor j; and $\Delta_{\mathrm{EE}}^{(s)}$ is as defined in (5). It is clear that the transformed sample $\tilde{\mathbf{d}}_l^{(s)} = \left(\tilde{d}_{l;1}^{(s)}, \tilde{d}_{l;2}^{(s)}, \ldots, \tilde{d}_{l;n}^{(s)}\right)$ has its mean equal to $\Delta_{\mathrm{EE}}^{(s)}$ for $s = 0,1$ and its variance being fixed at σ_l^2.

When testing the significance of the interaction effect of factor j via the lth hypothesis test with $l = 2j$ for $j = 1,2,\ldots,k$, the following transformation equation is applied to obtain $\tilde{\mathbf{d}}_l^{(s)}$ for $s = 0$ and 1:

$$\tilde{d}_{l;i}^{(s)} = \frac{d_{j;i}}{\widehat{\sigma}_j(n)} \Delta_{\mathrm{IE}}^{(s)}, \quad i = 1,2,\ldots,n, \tag{20}$$

where $\widehat{\sigma}_j(n)$ denotes the sample standard deviation of the elementary effects in $\mathbf{d}_j$, and $\Delta_{\mathrm{IE}}^{(s)}$ is as defined in (7). We see that the transformed sample $\tilde{\mathbf{d}}_l^{(s)}$ has its mean equal to zero with its variance equal to $\Delta_{\mathrm{IE}}^{(s)}$ for $s = 0, 1$.

With the transformed sample $\tilde{\mathbf{d}}_l^{(s)} = \left(\tilde{d}_{l;1}^{(s)}, \ldots, \tilde{d}_{l;n}^{(s)}\right)$ for the lth effect, we obtain an EPDF using (18), say $\check{f}_l(\cdot|\tilde{\mathbf{d}}_l^{(s)})$ for $s = 0, 1$. It is clear that $\check{f}_l(\cdot|\tilde{\mathbf{d}}_l^{(s)})$ satisfies the corresponding hypothesis of interest specified in either (6) or (7). Finally, the kernel-based SPRT test statistic for the lth test can be constructed as follows:

$$\widehat{\Lambda}_l(n) = \sum_{i=1}^{n} \left(\log \check{f}_l \left(d_{j,i} | \tilde{\mathbf{d}}_l^{(1)} \right) - \log \check{f}_l \left(d_{j,i} | \tilde{\mathbf{d}}_l^{(0)} \right) \right) . \tag{21}$$

Compared to the original test statistic $\Lambda_l(n)$ that is available only if the exact distribution is known, the calculation of $\widehat{\Lambda}_l(n)$ only requires a sample of elementary effects for factor $\lceil l/2 \rceil$ to obtain two kernel-based EPDFs that correspond to H_l and G_l, respectively.

3.5 Specification of Stopping Rules

In this paper we adopt a common pair of stopping boundaries for testing the $2k$ effects, which can be regarded as a type of *one-shot* boundary stopping strategy. Specifically, let $a_l \equiv a$ and $b_l \equiv b$ be the upper and lower stopping boundaries for $l = 1, 2, \ldots, 2k$. The stopping time of the lth hypothesis test in (10) reduces to

$$T_l = \inf \left\{ n \geq n_0 : \widehat{\Lambda}_l(n) \notin (b, a) \right\}, \tag{22}$$

where $\widehat{\Lambda}_l(n)$ is as given in (21). Subsequently, the stopping time of the entire procedure becomes

$$T = \inf \left\{ n \geq n_0 : \bigcap_{l=1}^{2k} \{ \widehat{\Lambda}_l(n) \notin (b, a) \} \right\}. \tag{23}$$

Following the discussion given in Subsection 3.2, an immediate choice for the upper and lower boundaries can be given as

$$a = -\log \left(\frac{\alpha}{2k} \right) \text{ and } b = \log \left(\frac{\beta}{2k} \right), \ l = 1, 2, \ldots, 2k, \tag{24}$$

where we set $\alpha_l = \alpha/2k$ and $\beta_l = \beta/2k$ for $l = 1, 2, \ldots, 2k$; and α and β are respectively the desired Type I and Type II familywise error levels specified in (8). It is easy to see that the two boundaries given by (24) are two horizontal lines in parallel with the horizontal axis that denotes the sample size n. The greater the values of α and β are, the narrower the continue-sampling region of the lth test becomes and the faster the test terminates.

4 NUMERICAL EVALUATION

In this section we demonstrate the performance of CMM on a factor screening problem, which has also been studied by Morris (1991) and Pujol (2009). The simulation output at a given factor combination $\mathbf{x} = (x_1, x_2, \ldots, x_{20})^\top \in [0, 1]^{20}$ is generated by

$$\mathscr{Y} = \beta_0 + \sum_{j=1}^{20} \beta_j w_j + \sum_{i<j}^{20} \beta_{i,j} w_i w_j + \sum_{\ell<i<j}^{20} \beta_{\ell,i,j} w_\ell w_i w_j + \sum_{s<\ell<i<j}^{20} \beta_{s,\ell,i,j} w_s w_\ell w_i w_j, \tag{25}$$

where $w_j \in [-1, 1]$ is transformed from $x_j \in [0, 1]$, according to the following two transformations: (1) the linear transformation $w_j = 2(x_j - 0.5)$, and (2) the nonlinear transformation $w_j = 2(1.1x_j/(x_j + 0.1) - 0.5)$, for $j = 1, 2, \ldots, 20$. The nonlinear transformation is applied to factors $j \in \{3, 5, 7\}$, and the linear

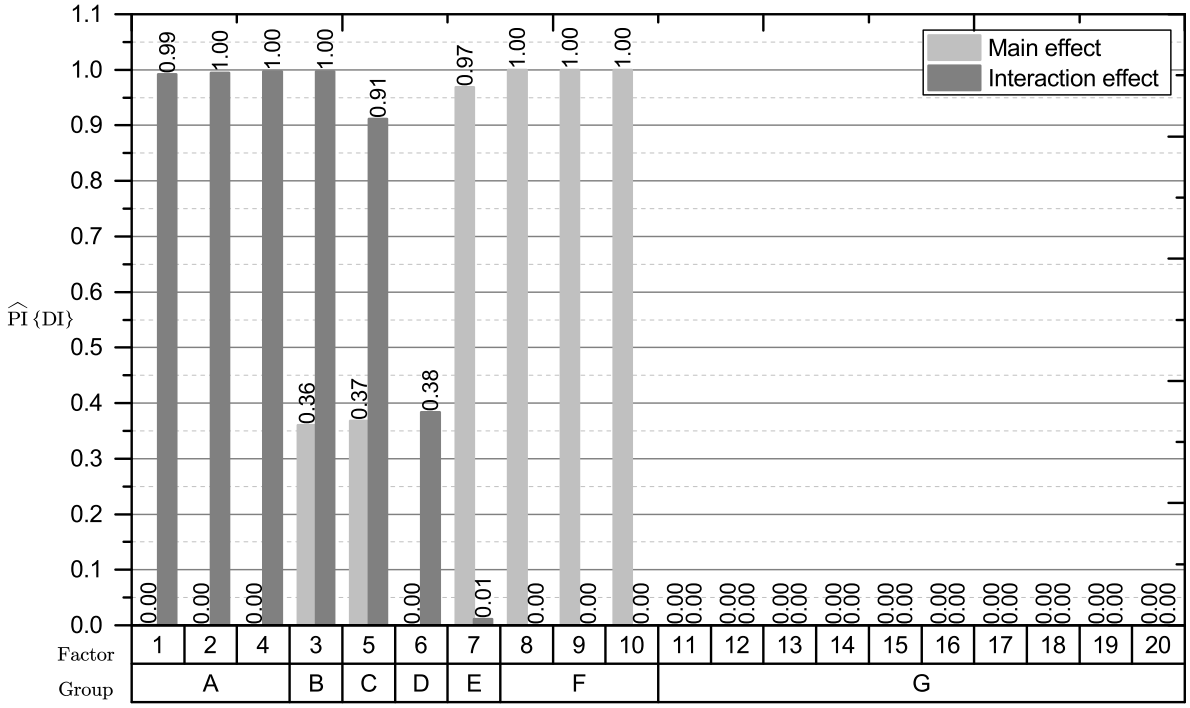

Figure 1: The resulting $\widehat{\mathrm{P}}\{\mathrm{DI}\}$ obtained by CMM for testing significance of the main and interaction effects of the 20 factors.

transformation is applied to the remaining 17 factors. We note that the value of w_j is uniformly distributed in $[-1,1]$ if the linear transformation is applied to x_j, whereas it is more likely to be in $[0,1]$ if the nonlinear transformation is applied. The coefficients in (25) are specified as follows. Regarding the first and second-order coefficients, $\beta_j = 20$ for $j \in \{1,2,\ldots,10\}$ and $\beta_{i,j} = 15$ for $i,j \in \{1,2,\ldots,6\}$; the remaining β_j's and $\beta_{i,j}$'s are independently sampled from standard normal distribution $\mathcal{N}(0,1)$. With respect to the third and fourth-order coefficients, $\beta_{\ell,i,j} = -10$ for $\ell,i,j \in \{1,2,\ldots,5\}$ and $\beta_{s,\ell,i,j} = 5$ for $s,\ell,i,j \in \{1,2,3,4\}$; the remaining $\beta_{\ell,i,j}$'s, $\beta_{s,\ell,i,j}$'s and β_0 are all set to zeros.

We assess the efficacy and efficiency of CMM for detecting important main and interaction effects associated with the 20 factors. To implement the sequential procedure of CMM, we use an initial sample size $n_0 = 20$ for all factors, and set the target Type I and Type II familywise error levels respectively at $\alpha = \beta = 0.1$. The threshold parameters $\Delta_{\mathrm{EE}}^{(0)}$, $\Delta_{\mathrm{EE}}^{(1)}$, $\Delta_{\mathrm{IE}}^{(0)}$, and $\Delta_{\mathrm{IE}}^{(1)}$ are set to 20, 30, 40, and 60, respectively. The entire procedure of CMM is applied for 1000 independent macro-replications, and the efficacy of CMM is evaluated by the fraction of times a given effect is declared significant, denoted by $\widehat{\mathrm{P}}\{\mathrm{DI}\}$, which is a commonly used measure in factor screening (Wan et al. 2010; Shi et al. 2014; Shi et al. 2016):

$$\widehat{\mathrm{P}}\{\mathrm{DI}\} = \frac{\#(\text{an effect is declared significant})}{\#\text{macro-replications}}.$$

The $\widehat{\mathrm{P}}\{\mathrm{DI}\}$'s obtained by CMM for testing the main and interaction effects of all 20 factors are shown in Figure 1. For each factor, two vertical bars are given that show the values of $\widehat{\mathrm{P}}\{\mathrm{DI}\}$ obtained for testing the significance of its corresponding main and interaction effects. We note that at the bottom of Figure 1 the 20 factors are also classified into 7 different groups "A–G" according to the coefficients associated with the terms involving each factor given in (25) and the transformation equations applied; see Shi et al. (2016) for details on how the groups are derived. The following observations can be made from Figure 1. First, the smaller μ_j (respectively, σ_j) is as compared to $\Delta_{\mathrm{EE}}^{(0)}$ (resp. $\Delta_{\mathrm{IE}}^{(0)}$), the closer the resulting $\widehat{\mathrm{P}}\{\mathrm{DI}\}$ for factor

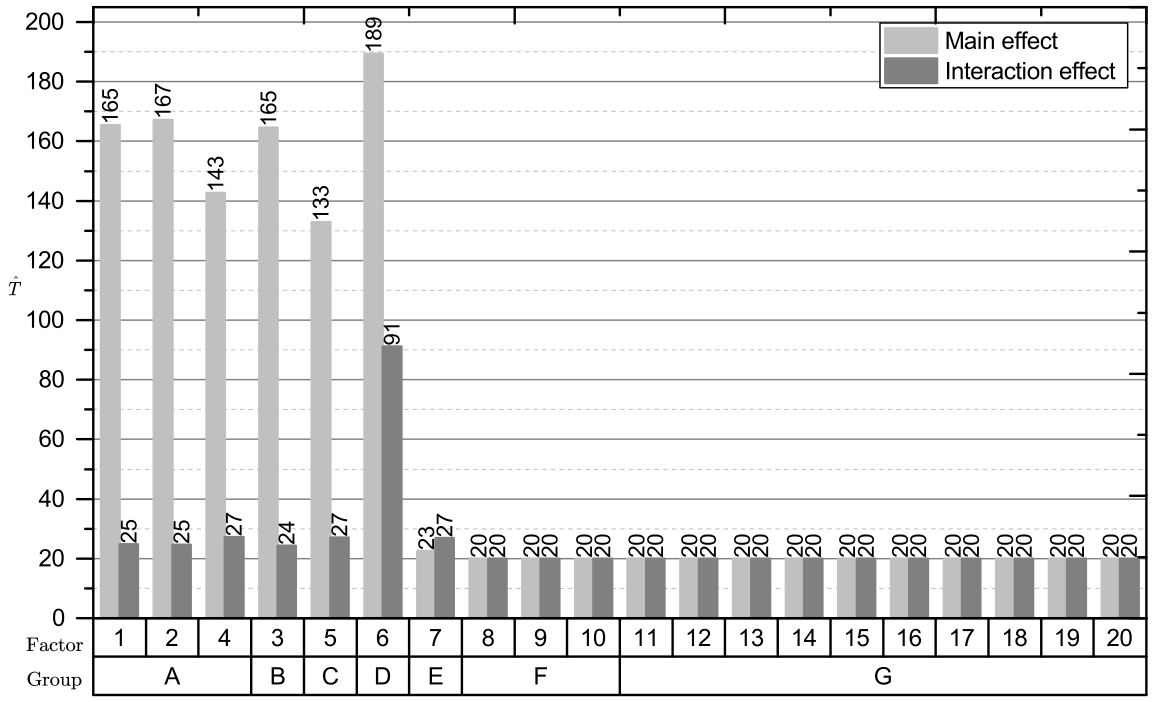

Figure 2: Comparison of average sample sizes used by CMM for testing significance of the main and interaction effects of the 20 factors.

j is to zero. Second, the greater μ_j (respectively, σ_j) is when compared to $\Delta_{\text{EE}}^{(1)}$ (resp. $\Delta_{\text{IE}}^{(1)}$), the closer the resulting $\widehat{\text{P}}\{\text{DI}\}$ for factor j gets to one. Third, for factors with $\mu_j \in (\Delta_{\text{EE}}^{(0)}, \Delta_{\text{EE}}^{(1)})$ or $\sigma_j \in (\Delta_{\text{IE}}^{(0)}, \Delta_{\text{IE}}^{(1)})$, the resulting $\widehat{\text{P}}\{\text{DI}\}$ takes reasonable values in $[0,1]$. Lastly, we note that CMM obtains almost identical $\widehat{\text{P}}\{\text{DI}\}$ for factors within the same group. Therefore, we conclude that CMM provides a desired statistical performance guarantee for testing the significance of the main and interaction effects of each individual factor.

The computational efficiency of CMM is quantified by the average sample size of elementary effects used by the sequential procedure for testing a given effect across the 1000 macro-replications. The respective sample sizes used for testing the significance of the main and interaction effects of each factor are shown in Figure 2. We observe that CMM adapts the sample sizes used for the 20 factors according to the magnitudes of their respective main and interaction effects. In particular, for those factors with a high σ_j and μ_j being close to $\Delta_{\text{EE}}^{(0)}$ or $\Delta_{\text{EE}}^{(1)}$, CMM typically uses a large sample size for testing the significance of the corresponding factor. It is intuitively clear that such an adaptive sampling strategy of CMM is more efficient than the default equal budget allocation rule adopted by standard MM.

5 CONCLUSIONS

In this paper, we propose the controlled Morris method (CMM) for factor screening that acts in a sequential manner to keep the computational effort down to a minimum. The SPRT-based multiple testing procedure adopted enables CMM to identify the factors with significant main and/or interaction effects while controlling Type I and Type II familywise error rates at desired levels. Though CMM is proposed in the context of factor screening, its distribution-free SPRT-based multiple testing procedure can be broadly applied to various settings beyond factor screening. Future research topics include extending CMM for factor screening in the stochastic simulation setting and enhancing the computational efficiency achieved by the sequential multiple testing procedure.

ACKNOWLEDGMENT

The work of Wen Shi was partially supported by the National Natural Sciences Foundation of China under Grants No. 71402048, 71372134, and 71671060, and the China Postdoctoral Science Foundation Funded Project No. 2015M582228. The work of Xi Chen was partially supported by the Virginia Tech ICTAS Junior Faculty Award.

REFERENCES

Ankenman, B. E., R. C. H. Cheng, and S. M. Lewis. 2014. "Screening for Dispersion Effects by Sequential Bifurcation". *ACM Transactions on Modeling and Computer Simulation* 100 (100): 1–27.

Antoniak, C., and G. Dillard. 1968. "A Distribution-free Sequential Probability-ratio Test for Multiple-resolution-element Radars". *IEEE Transactions on Information Theory* 14 (6): 822–825.

Bartroff, J., and J. Song. 2014. "Equential Tests of Multiple Hypotheses Controlling Type I and II Familywise Error Rates". *Journal of Statistical Planning and Inference* 153:100–114.

Bettonvil, B., and J. P. C. Kleijnen. 1997. "Searching for Important Factors in Simulation Models with Many Factors: Sequential Bifurcation". *European Journal of Operational Research* 96 (1): 180–194.

Boukouvalas, A., J. P. Gosling, and H. Maruri-Aguilar. 2014. "An Efficient Screening Method for Computer Experiments". *Technometrics* 56 (4): 422–431.

Campolongo, F., and R. Braddock. 1999. "The Use of Graph Theory in the Sensitivity Analysis of the Model Output: A Second Order Screening Method". *Reliability Engineering & System Safety* 64 (1): 1–12.

Campolongo, F., J. Cariboni, and A. Saltelli. 2007. "An Effective Screening Design for Sensitivity Analysis of Large Models". *Environmental Modelling & Software* 22 (10): 1509–1518.

Cropp, R. A., and R. D. Braddock. 2002. "The New Morris Method: An Efficient Second-order Screening Method". *Reliability Engineering & System Safety* 78 (1): 77–83.

De, S. K., and M. Baron. 2012a. "Sequential Bonferroni Methods for Multiple Hypothesis Testing With Strong Control of Family-wise Error Rates I and II". *Sequential Analysis* 31 (2): 238–262.

De, S. K., and M. Baron. 2012b. "Step-up and Step-down Methods for Testing Multiple Hypotheses in Sequential Experiments". *Journal of Statistical Planning and Inference* 142 (7): 2059–2070.

Deng, Z., F.-L. Chung, and S. Wang. 2008. "FRSDE: Fast Reduced Set Density Estimator Using Minimal Enclosing Ball Approximation". *Pattern Recognition* 41 (4): 1363–1372.

Dykstra, R. 2005. *Kullback-Leibler Information*. John Wiley & Sons, Ltd.

Fédou, J. M., and M. J. Rendas. 2015. "Extending Morris Method: Identification of the Interaction Graph Using Cycle-equitable Designs". *Journal of Statistical Computation and Simulation* 85 (7): 1398–1419.

Han, B., D. Comaniciu, Y. Zhu, and L. S. Davis. 2008, July. "Sequential Kernel Density Approximation and Its Application to Real-time Visual Tracking". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (7): 1186–1197.

Härdle, W. 1990. *Applied Nonparametric Regression*. Cambridge University Press.

Kristan, M., A. Leonardis, and D. Skočaj. 2011. "Multivariate Online Kernel Density Estimation with Gaussian Kernels". *Pattern Recognition* 44 (10C11): 2630–2642.

Kristan, M., D. Skoaj, and A. Leonardis. 2010. "Online Kernel Density Estimation for Interactive Learning". *Image and Vision Computing* 28 (7): 1106–1116.

López-Rubio, E., and J. M. O. de Lazcano-Lobato. 2008. "Soft Clustering for Nonparametric Probability Density Function Estimation". *Pattern Recognition Letters* 29 (16): 2085–2091.

Morris, M. D. 1991. "Factorial Sampling Plans for Preliminary Computational Experiments". *Technometrics* 33 (2): 161–174.

Parpas, P., B. Ustun, M. Webster, and Q. K. Tran. 2015. "Importance Sampling in Stochastic Programming: a Markov Chain Monte Carlo Approach". *INFORMS Journal on Computing* 27 (2): 358–377.

Pujol, G. 2009. "Simplex-based Screening Designs for Estimating Metamodels". *Reliability Engineering and System Safety* 94:1156–1160.

Shi, W., X. Chen, and J. Shang. 2016. "An Efficient Morris Method-based Framework for Simulation Factor Screening". Technical Report, Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, USA.

Shi, W., J. P. C. Kleijnen, and Z. X. Liu. 2014. "Factor Screening for Simulation With Multiple Responses: Sequential Bifurcation". *European Journal of Operational Research* 237 (1): 136–147.

Wald, A. 1992. *Sequential Tests of Statistical Hypotheses*, 256–298. New York, NY: Springer New York.

Wan, H., B. E. Ankenman, and B. L. Nelson. 2010. "Improving the Efficiency and Efficacy of Controlled Sequential Bifurcation for Simulation Factor Screening". *INFORMS Journal on Computing* 22 (3): 482–492.

Wang, W., and H. Wan. 2014. "Sequential Procedures for Multiple Responses Factor Screening". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 745–756. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Woods, D. C., and S. M. Lewis. 2016. *Design of Experiments for Screening*, 1–43. Cham: Springer International Publishing.

Yu, C., and B. Su. 2004. "A Non-parametric Sequential Rank-sum Probability Ratio Test Method for Binary Hypothesis Testing". *Signal Processing* 84 (7): 1267–1272.

Zhang, G., and Y. Wang. 2014. "A Gaussian Kernel-based Approach for Modeling Vehicle Headway Distributions". *Transportation Science* 48 (2): 206–216.

AUTHOR BIOGRAPHIES

WEN SHI is an Associate Professor in School of Logistics and Engineering Management at Hubei University of Economics. His research focuses on simulation modeling, simulation experiments design and analysis, and supply chain management. His email address is shi3wen@163.com.

XI CHEN is an Assistant Professor in the Grado Department of Industrial and Systems Engineering at Virginia Tech. Her research interests include stochastic modeling and simulation, applied probability and statistics, computer experiment design and analysis, and simulation optimization. Her email address is xchen6@vt.edu and her web page is https://sites.google.com/vt.edu/xi-chen-ise/home.