Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

SEQUENTIAL PROCEDURES FOR MULTIPLE RESPONSES FACTOR SCREENING

Wenyu Wang
Hong Wan

School of Industrial Engineering
Purdue University
315 North Grant Street
West Lafayette, IN 47907, USA

ABSTRACT

This paper considers the factor screening problem with multiple responses for simulation experiments. The objective is to identify important factors with controlled Family-Wise Error Rate. We assume a multiple-response first-order linear model, the responses follow a multivariate normal distribution, and estimated effect coefficients also follow multivariate normal distribution. Two likelihood ratio based procedures, Sum Intersection Procedure (SUMIP) and Sort Intersection Procedure (SORTIP), are proposed and verified. Numerical studies are provided to demonstrate the validity and efficiency of our proposed procedures.

1 INTRODUCTION

Simulation models for complex systems typically involve hundreds to thousands of factors. According to Pareto or sparsity-of-effects principle, in many cases only a few factors among many are responsible for most of the response variation (Myers, Anderson-Cook, and Montgomery 2009). Factor screening experiments are designed to identity a small subset of important factors efficiently so that the later effort can be focused on them, therefore significantly save the overall experimental effort.

There has been considerable research in this area. Procedures including one-factor-at-a-time designs (Saltelli, Tarantola, and Campolongo 2000), edge designs (Elster and Neumaier 1995), and the Trocine screening procedure (Trocine and Malone 2000) are designed for stochastic simulation experiments. However, these procedures are based on homogeneous variance assumption and, more importantly, fail to provide the error control. Later, procedures including Controlled Sequential Bifurcation (Wan, Ankenman, and Nelson 2006), Two-stage Controlled Fractional Factorial Screening (TCFF) (Wan and Ankenman 2007), Controlled Sequential Factorial Design (Shen and Wan 2009), and hybrid method of CSB and CSFD (Shen, Wan, and Sanchez 2010) were proposed to address error controls on both Type I error and Type II error, and relax the homogeneous variance requirement. The interested reader should refer to Kleijnen et al. (2005) for reviews.

Most of the previous literature as discussed above focused on the single response model. In practice, however, there are usually multiple responses of interest that can be observed simultaneously in one experiment. For example, risk and return are two responses of a portfolio, and finding factors that significantly contribute to these two responses is of great importance. The main challenge for multiple response factor screening is to achieve desired error control with efficiency. Note that the multiple responses add another layer of complexity to the error control. Lee, Chew, and Teng (2007) studied how to allocate computation resources efficiently to identify "Pareto Set" of designs for a multiple objective Ranking and Selection Problem. The concept of "Pareto Set" inspires our definition of important/unimportant factors in this paper. Multiple responses factor screening problem has been firstly studied by Shi, Kleijnen, and Liu (2014). They proposed Multiple Sequential Bifurcation (MSB) to identify important factors. MSB generalizes CSB (Wan, Ankenman, and Nelson 2006) and adopts sequential bifurcation to classify a group or an individual factor as important or unimportant. However, MSB is conservative and restricted since it controls error rate and power through Bonferroni procedures and fails to incorporate covariance information among responses.

In this paper, we propose two sequential hypothesis testing procedures, SUMIP and SORTIP, for multi-response factor screening. Both procedures control the Family-Wise Error Rates (FWERs), instead of individual Type I error rate and power for each factor. We consider two types of FWER, Family-Wise Error Rate I and Family-Wise Error Rate II, which are defined as follows:

Definition 1 Let V denote the number of true null hypotheses being rejected and T the number of false null hypotheses being accepted. The Family-Wise Error Rate I (FWERI) is defined as the probability to reject at least one true null hypothesis; the Family-Wise Error Rate II (FWERII) is the probability of accepting at least one false null hypothesis.

$$\text{FWERI} = Prob\,(V \geq 1)$$
$$\text{FWERII} = Prob\,(T \geq 1)$$

Note that FWERI and FWERII are defined independent of the number of individual tests. Find more discussions on FWERI and FWERII in Lee Lee (2004).

Both SUMIP and SORTIP are Sequential Probability Ratio Test (SPRT) procedures. SPRT is a sequential likelihood-ratio testing procedure that varies with the specifics of hypotheses and assumed distributions. From Neyman-Pearson lemma, among all tests with the same significance level, likelihood ratio test has the greatest power. Moreover, the sequential nature of SPRT allows it to achieve the same power more efficiently than fixed number of samples (Wald and Wolfowitz 1948). SPRT first chooses a pair of constants A and B with $0 < A < 1 < B < \infty$ as thresholds of likelihood ratio. After each observation, SPRT will calculate Λ, the likelihood ratio of hypothesis H_1 against H_0, and choose H_0 if Λ is smaller than A, H_1 if Λ is greater than B, or take one more observation. The smaller the A, the smaller the value of Λ needed to accept the null hypothesis, and then the greater the power; and the greater the B, the larger the value of Λ needed to reject the null hypothesis and then the smaller the Type I error.

There are two phases in SUMIP and SORTIP. Phase 1 is to calculate Likelihood Ratio (LR) for each factor; phase 2 is to check whether LRs of all factors are sufficient to categorize factors into important and unimportant sets within permitted FWERI and FWERII. For Phase 1, we derive formula and algorithm to calculate likelihood ratio for each factor. In this problem, individual test on each factor takes the forms of multivariate normal union-intersection tests. Literatures provide two other test methods. First method is to approximate the test as Hotelling t test or χ^2 test. These tests are substantially less powerful than likelihood ratio test (LRT) for testing multivariate normal mean vector $H_0: \ \mu = 0$ against $H_1: \ \mu \neq 0$ according to Neymann-Pearson Lemma (Wald and Wolfowitz 1948). Moreover, χ^2 or Hotelling t tests are bias for one-sided test (Perlman and Wu 2006). Another choice is to decompose the testing problem into a collection of sub-problems and treat them independently. Reitmeir and Wassmer (1996), Perlman and Wu (2004), and Logan () discussed procedures in this fashion. However, their procedures neither fit sequential manner nor take advantage of covariance information. In this paper, we provide the exact likelihood ratio and prove that maximum likelihood estimator (MLE) is the solution of a quadratic programming problem. Also, our method is robust under various covariance structure as demonstrated in our numerical evaluation.

For Phase 2, we propose two efficient and powerful stopping criteria to control FWERI and FWERII. The complexity of such schemes lies in the number of possible hypotheses combinations. For a single test, there are only 2 possibilities H_0 and H_1. However, for familywise test that formed by I individual tests, the decision schemes should be able to make decisions among 2^I possibilities. The common practice is to apply Bonferroni method among these I tests, which is intuitive but conservative. So far, the most efficient procedure is the Holm Intersection Procedure (HIP) proposed by De and Baron (2012). HIP generalizes

Holm's procedure, which is designed to control only FWERI, to both FWERI and FWERII in a sequential manner. In our procedures, we propose weaker stopping criteria than HIP. Numerical results show that our proposed methods achieve the same power as HIP with only approximately half of the sample size.

In our numerical evaluation, we use factorial design to estimate factor coefficients. Thus, one sample is equivalent to one replicate of the factorial design. It is worth noting here that SUMIP and SORTIP can be used for any designs, as long as the estimated coefficient vector of each factor follows a multivariate normal distribution.

The rest of this paper is organized as follows: Section 2 states the problem details and assumptions. Section 3 presents the two proposed procedures. Numerical evaluate are given in Section 4. We conclude the paper with discussions of future research in section 5.

2 PROBLEM STATEMENT

In this paper, we assume a first-order model for all responses,

$$Y_j = \beta_{0j} + \sum_{i=1}^{I} \beta_{ij} x_i + \varepsilon_j,\ \varepsilon_j \sim N(0, \sigma_j^2),\ \sigma_j \propto Y_j,\ j = 1, 2, ...J \tag{1}$$

which could also be written in matrix form as

$$Y = X'\beta + \varepsilon \tag{2}$$

where $Y = [Y_j]_J$, $X = [x_i]_I$, $\varepsilon = [\varepsilon_j]_J$ and $\beta = [\beta_{ij}]_{I \times J}$.

Here, Y_j is the output of the j^{th} response, and β_{ij} represents the main effect of the i^{th} factor on the j^{th} response. Quantitative inputs, x_i, are normalized such that they have only two levels which are arbitrarily denoted as -1 or 1. ε are the metamodel residuals with mean 0. Furthermore, we assume ε_j for $j = 1, 2, ..., J$ can be correlated and their scale may depend on the size of the output. In other words, we assume (1) unknown covariance structure among different responses, and (2) heterogeneous variance condition.

To carry on, we need to define "important factor" in multiple responses setting. We use definitions from Lee, Chew, and Teng (2007) and Shi, Kleijnen, and Liu (2014), where important factors are defined as factors which have important effect on at least one response. Let Δ_{0j} be the threshold for important effect on the jth response, and Δ_{1j} the threshold for critical effect on the jth response. Thus,

Definition 2 The ith factor β_i is **unimportant** if $|\beta_{ij}| \le \Delta_{0j}$ for all $j = 1, 2, ..., J$, **important** if there exists at least one j such that $|\beta_{ij}| \ge \Delta_{0j}$, and **critical** if there exists at least one j such that $|\beta_{ij}| \ge \Delta_{1j}$. Specifically,

$$\Theta_c = \left\{ \beta_i | \max_{j \in \{1,2,..,J\}} \left(|\beta_{ij}| - \Delta_{1j} \right) \ge 0,\ i \in \{1, 2, ..I\} \right\}$$

$$\Theta_{imp} = \left\{ \beta_i | \max_{j \in \{1,2,..,J\}} \left(|\beta_{ij}| - \Delta_{0j} \right) \ge 0,\ i \in \{1, 2, ..I\} \right\}$$

$$\Theta_0 = \left\{ \beta_i | \max_{j \in \{1,2,..,J\}} \left(|\beta_{ij}| - \Delta_{0j} \right) \le 0,\ i \in \{1, 2, ..I\} \right\}$$

With this definition, our goal is, under proper error control, to screen out factors within Θ_0 and claim them unimportant, and to find out factors in Θ_c, which have great impacts on responses, and claim them important. For factors within the "indifference zone", which means factor is in Θ_{imp} but not in Θ_c, although they are important, it is neither possible nor our intention to control the power for these factors.

For clarity, we call the single test for each factor the **elementwise test**, and the I elementwise tests form the **familywise test**.

In this paper, elementwise test is defined as follows:

$$H_0^i:\ \beta_i \in \Theta_0 \quad H_1^i:\ \beta_i \in \Theta_{imp} \quad \text{for } i = 1, ..., I \tag{3}$$

This elementwise test can be shown in the Union-Intersection form,

$$H_0^i = \bigcap_{j=1}^{J} H_o^{ij} \quad H_1^i = \bigcup_{j=1}^{J} H_1^{ij}$$

where

$$H_0^{ij}:\ \beta_{ij} < \Delta_{0j} \quad H_1^{ij}:\ \beta_{ij} \geq \Delta_{0j} \text{ for } j = 1, ..., J \tag{4}$$

Familywise test includes I elementwise tests. As discussed earlier, our intention is to control FWERI and FWERII, where FWERI is the probability of identifying at least one unimportant factor as important and FWERII is the probability of classifying at least one critical factor as unimportant.

3 PROCEDURES

Procedures proposed in this paper contain two phases. Phase 1 is to calculate Likelihood Ratio (LR) for each factor, whose effects follow multivariate normal distribution; phase 2 is to aggregate LRs of all factors and decide whether the aggregated LR is sufficient to make statistic inference within upperbounds of FWERI and FWERII. This section will discuss these two phases and state the algorithms.

3.1 Phase 1: Likelihood Ratio for Multivariate Normal Union-Intersection Test

Recall the test for factor i,

$$H_0^i:\ \beta_i \in \Theta_0 \quad H_1^i:\ \beta_i \in \Theta_{imp} \tag{5}$$

From experiments, we observe $\{Z_{i1},\ Z_{i2}, ...Z_{iK}\}$ as K estimators of β_i, where Z_{ik} is of dimension $J \times 1$ and from distribution $N(\beta_i, \Sigma_i)$ with unknown mean β_i and covariance matrix Σ_i. If $K \geq J+1$, then $S_i = \sum_{k=1}^{K} (Z_{ik} - \bar{Z}_i)(Z_{ik} - \bar{Z}_i)'$ is positive definite with probability one.

Assuming $\{Z_{i1},\ Z_{i2}, ...Z_{iK}\}$ are independent. $\bar{Z}_i$ and S_i are then sufficient statistics for β_i and Σ, and $\sqrt{K}\bar{Z}_i \sim N(\beta_i, \Sigma_i)$, $S_i \sim \text{Wishart}(K-1, \Sigma_i)$, and $\bar{Z}_i$ and S_i are independent. Based on these assumptions, we can deduct the likelihood function and likelihood ratio for (5), which are given in Theorem 1.

Theorem 1 Likelihood ratio for (5) is

$$\Lambda_i(\Theta_0, \Theta_c) = \frac{L_1}{L_0} = \frac{\left(1 + \|\bar{Z}_i - \beta_{mle}(\bar{Z}_i, S_i, \Theta_c)\|^2_{\frac{S_i}{K}}\right)^{-\frac{K}{2}}}{\left(1 + \|\bar{Z}_i - \beta_{mle}(\bar{Z}_i, S_i, \Theta_0)\|^2_{\frac{S_i}{K}}\right)^{-\frac{K}{2}}}$$

where $\|Z - \beta\|_S^2 = (\beta - Z)' S^{-1} (\beta - Z)$ and $\beta_{mle}(Z, S, \Theta)$ is the maximum likelihood estimator for β within the parameter space Θ,

$$\beta_{mle}(Z, S, \Theta) = \arg\min_{\beta \in \Theta} \|Z - \beta\|_S^2$$

For $\Theta = \Theta_0$, $\beta_{mle}(Z, S, \Theta_0)$ is the solution of the following quadratic programming problem with convex constraints,

$$\begin{aligned} &\min (\beta - Z)' S^{-1} (\beta - Z) \\ &\text{s.t.} \left|\beta_j\right| \leq \Delta_{0j} \text{ for } j = 1, 2, ..., J \end{aligned} \tag{6}$$

For $\Theta = \Theta_c$, $\beta_{mle}(Z,S,\Theta_c)$ is the solution of the following problem,

$$\begin{aligned} &\min (\beta - \bar{Z})' S^{-1} (\beta - \bar{Z}) \\ &\text{s.t.} \left|\beta_j\right| \geq \Delta_{1j} \text{ for at least one } j,\ j = 1,2,...,J \end{aligned} \tag{7}$$

Note that this is no longer a convex quadratic programming problem. We prove that it can be solved as a convex problem as demonstrated in Theorem 2.

Theorem 2 If $Z \in \Theta_c$, $\beta_{mle}(Z,S,\Theta_c) = Z$;

If $Z \notin \Theta_c$, $\beta_{mle}(Z,S,\Theta_c)$ is the solution of the following quadratic programming problem,

$$\begin{aligned} &\min (\beta - Z)' S^{-1} (\beta - Z) \\ &\text{s.t. } \beta_{j^\star} = \text{sign}(Z_{j^\star})\Delta_{1j^\star} \end{aligned} \tag{8}$$

where $j^\star = \arg\min_{j=1,...,J} \frac{\left(\Delta_{1j} - z_j\right)^2}{s_j}$ with $s_j > 0$ as the j^{th} diagonal element of S.

Thus, both $\beta_{mle}(Z,S,\Theta_0)$ and $\beta_{mle}(Z,S,\Theta_c)$ could be modeled as quadratic optimization problems with linear constraints and, thus, be solved easily. (Functions, such as `quadprog` in Matlab and `quadprog` in R, are for this type of problem.)

3.2 Phase 2: Sum Intersection Scheme and Sort Intersection Scheme

For familywise test, we could choose whether to make decision for elementwise tests simultaneously or separately, which correspond to two types of stopping rules. One is the intersection rule, which makes simultaneously inferences; the other one is the maximum rule, which treats I elementwise tests separately. The intersection rule is less conservative since it dynamically allocate the "allowed errors" according to the likelihood ratios, instead of assigning the same amount of errors to each factor (De and Baron 2012).

3.2.1 Intersection Rule

Let A_i and B_i ($0 < A_i < 1 < B_i < \infty$) be the pair of thresholds for Likelihood Ratio Test of the i^{th} elementwise test. The intersection rule states that if all elementwise tests reach decision regions, which means that the likelihood ratio $\Lambda^i \notin (A_i, B_i)$ for all $i \in \{1,2,...I\}$, then stop sampling and choose H_{0i} if Λ^i is smaller than A_i or H_{1i} if Λ_i is greater than B_i; otherwise take one more observation for all elementwise tests. Therefore, the intersection rule stops at the first time when $\Lambda_n^i \notin (A_i, B_i)$ for all $i = 1,2,3,...I$.

$$N_{int} = \inf\left\{ n : \bigcap_{i=1}^{I} \Lambda_n^i \notin (A_i, B_i) \right\}$$

Intersection rule is a proper stopping time since it stops in finite time with probability 1 (De and Baron 2012). The following corollary reveals the connection between thresholds, A_i and B_i, and Type I error and Type II error for the intersection rule.

Corollary 3 (De and Baron, 2012) Let τ be any stopping time satisfying

$$Prob\left(\Lambda_\tau^i \in (A_i, B_i)\right) = 0 \ for\ A_i < 1,\ B_i > 1\ for\ i \in \{1,2,...,I\}$$

with the decision rule of rejecting H_0^i if and only if $\Lambda_\tau^i \geq A_i$. For such a test

$$Prob\left(\text{Type I on } i^{th} \text{ test}\right) = Prob\left(\Lambda_\tau^i > B_i | H_{0i} \text{ is true}\right) \le B_i^{-1}$$
$$Prob\left(\text{Type II on } i^{th} \text{ test}\right) = Prob\left(\Lambda_\tau^i < A_i | H_{1i} \text{ is true}\right) \le A_i$$

Based on Corollary 3, we propose two schemes, Sum Intersection Scheme and Sort Intersection Scheme, to connect thresholds, A_i and B_i, with Family Wise Error Rate.

3.2.2 Sum Intersection Scheme

We propose Sum Intersection Scheme with the intersection stopping time and boundaries as

$$N_{int} = \inf\left\{ n : \bigcap_{i=1}^{I} \Lambda_n^i \notin (A_i, B_i) \right\}, \ \sum_{i \in I_A} A_i \le \gamma, \ \sum_{i \in I_B} B_i^{-1} \le \alpha \tag{9}$$

where $I_A = \left\{ i : \Lambda_n^i \le 1 \right\}$, $I_B = \left\{ i : \Lambda_n^i > 1 \right\}$ and α and γ are the predefined FWERI and FWERII, respectively. I_A and I_B are determined after each iteration and hence may change at each step.

At N_{int}, factors within I_B are claimed important, while factors within I_A unimportant.

Theorem 4 The Sum Intersection Scheme strongly controls both FWERI and FWERII. That is,

$$FWERI = Prob\left(\text{At least 1 Type I error among I tests}\right) \le \alpha$$
$$FWERII = Prob\left(\text{At least 1 Type II error among I tests}\right) \le \gamma$$

The insight of this scheme is to apply Bonferroni procedure, and set $B_i = \infty$ for factors within I_A and $A_i = 0$ for factors within I_B.

3.2.3 Sort Intersection Scheme

We propose Sort Intersection Scheme with stopping time and boundaries as,

$$N_{int} = \inf\left\{ n : \left(\bigcap_{i \in I_A} \Lambda_n^{(i)} \notin (A_i, 1) \right) \bigcap \left(\bigcap_{j \in I_B} \Lambda_n^{(j)} \notin (1, B_j) \right) \right\}, \ A_i = \frac{\gamma}{|I_A| - i + 1}, \ B_j = \frac{j}{\alpha} \tag{10}$$

where α and γ be the upperbounds of FWERI and FWERII, and $\Lambda_n^{(i)}$ is the ith smallest likelihood ratios. After each sampling, factors are sorted in increasing order and then separated into two groups according to their likelihood ratios, $I_A = \left\{ i : \Lambda^i \le 1 \right\}$ and $I_B = \left\{ i : \Lambda^i > 1 \right\}$, with cardinalities $|I_A|$ and $|I_B|$. For factors within I_A, it is certain that we wouldn't commit Type I error. Thus, we only need to consider FWERI for factors within I_B. Same logic applies to factors within I_B.

We can prove that the Sort Intersection Scheme controls both FWERI and FWERII in the strong sense.

Theorem 5 The Sort Intersection Scheme strongly controls both FWERI and FWERII. That is,

$$FWERI = Prob\left(\text{At least 1 Type I error among I tests}\right) \le \alpha$$
$$FWERII = Prob\left(\text{At least 1 Type II error among I tests}\right) \le \gamma$$

3.3 Procedures

We now apply the aforementioned schemes into screening procedures. It is worth mentioning that the hypothesis testing procedures can be applied in screening problem with any experimental designs as long as the estimator of the factor effect follows multivariate normal distribution. In particular, we demonstrate these procedures with the fractional factorial design.

3.3.1 Sum Intersection Procedure

The first procedure, Sum Intersection Procedure (SUMIP for short), is a two-phase procedure.

Phase 1 calculates likelihood ratio for each elementwise test. Phase 2 applies Sum Intersection Scheme. This procedure calculates Likelihood Ratios (LRs) after sampling; then assigns factors into I_B if $LR > 1$ or I_A otherwise. If LRs in I_B and I_A satisfy (9), then we claim that factors within I_A are unimportant and I_B are important. If not, we collect one more set of samples. Notice that Y_k is the kth simulation result.

Algorithm 1 SUMIP Procedure

Step 0 Select a factorial design X for I factors, generate n_0 replications of observations. Let $n = n_0$.

Step 1

Likelihood Ratio Calculate Likelihood Ratio Λ_n^i for each factor $i \in \{1,2,...,I\}$ by Theorem 1 with

$$Z_k = \left(X'X\right)^{-1} X'Y_k,\ k = 1,2,...n_0$$
$$S_n = \sum_{k=1}^{n} (Z_k - \bar{Z})(Z_k - \bar{Z})'$$

Stopping Criteria Apply Sum Intersection Scheme

Divide factors into two set $I_A = \left\{i : \Lambda_n^i \leq 1\right\}$ and $I_B = \left\{i : \Lambda_n^i > 1\right\}$

If $\alpha \geq \sum_{i \in I_B} \left(\Lambda_n^i\right)^{-1}$ and $\gamma \geq \sum_{i \in I_A} \Lambda_n^i$, go to Step 2.

Otherwise, generate 1 more observation with X, and make $n = n+1$; back to Step 1.

Step 2 Claim factors within I_B are important, and factors within I_A are unimportant.

3.3.2 Sort Intersection Procedure

Algorithm 2 SORTIP Procedure

Step 0 Select a factorial design X for I factors, generate n_0 replications of observations. Let $n = n_0$.

Step 1

Stopping Criteria Calculate Likelihood Ratio Λ_n^i for each factor $i \in \{1,2,...,I\}$ by Theorem 1 with

$$Z_k = \left(X'X\right)^{-1} X'Y_k,\ k = 1,2,...n_0$$
$$S_n = \sum_{k=1}^{n} (Z_k - \bar{Z})(Z_k - \bar{Z})'$$

Stopping Criteria Apply Sort Intersection Scheme

Divide factors into two set $I_A = \left\{i : \Lambda_n^i \leq 1\right\}$ and $I_B = \left\{i : \Lambda_n^i > 1\right\}$ and sort $\Lambda_n^i, i \in I_A$ and $\Lambda_n^j, j \in I_B$ ascending.

If $\Lambda_n^{(i)} \notin (A_i, 1)$ where $A_i = \frac{\gamma}{I_A - i + 1}$, and $\Lambda_n^{(j)} \notin \left(1, B_j\right)$ where $B_j = \frac{j}{\alpha}$ for all $i \in I_A,\ j \in I_B$, then go step 2.

Otherwise, generate 1 more observation with X, and make $n = n+1$; back to step 1.

Step 2 Claim factors within I_B are important, and factors within I_A are unimportant.

This procedure, Sort Intersection Procedure (SORTIP), is based on Sort Intersection Scheme. It has the same structure as SUMIP except for stopping criteria. This procedure will run n_0 initial replications of the design at first, where $n_0 \geq J$; then split factors into two groups, $I_A = \left\{i : \Lambda^i \leq 1\right\}$ and $I_B = \left\{i : \Lambda^i > 1\right\}$. If likelihood ratios satisfy (10), then we claim factors within I_B as important and I_A as unimportant. If not, we conduct one more factorial design for all factors.

4 NUMERICAL EXAMPLES

In this section, we compare our proposed procedures, SORTIP and SUMIP, with HIP proposed by De and Baron (2012) and MSB (Shi, Kleijnen, and Liu 2014) to demonstrate the efficiency and validity of both SORTIP and SUMIP. MSB procedure controls Type I error rate and Type II error rate for elementwise test. Thus, we compare Likelihood ratio testing procedure used as Phase 1 in SUMIP and SORTIP with MSB. HIP differs from SUMIP and SORTIP in stopping criteria. We run all three procedures for factor screening in the second experiment to compare their performances and conclude both SUMIP and SORTIP outperform HIP.

Numerical simulations use common random numbers across different methods to compare these algorithms.

4.1 Single-Factor Multiple-Responses Study

This experiment is to show the performance of the proposed Likelihood Ratio Test, which is phase 1 of SORTIP and SUMIP, with MSB. Consider a factor screening problem with a single factor and multiple responses. Parameters are set as follows:

Table 1: Simulation experiment parameter of single factor screening.

Parameter	Value	Meaning
J	Specified below	Number of Responses
Δ_0	2	Threshold of Important Factors
Δ_1	4	Threshold of Critial Factors
α	0.05	Upperbound of Type I Error
γ	0.05	Upperbound of Type II Error
β	Specified below	Responses size
Σ	Specified below	Covariance matrix
n_0	6	Initial sample size

In all cases, the presented results are the averages of 1000 independent macro replications.

For this example, we take three covariance matrix forms to represent independent case, positive dependent case, and nested case, respectively. For nested case, there are both positive and negative correlations among responses.

$$\Sigma_1 = \begin{bmatrix} 4 & 0 & & 0 \\ 0 & 4 & & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & 4 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 4 & 1 & & 1 \\ 1 & 4 & & 1 \\ & & \ddots & \vdots \\ 1 & 1 & \cdots & 4 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 4 & (-1)^2 & & (-1)^J \\ (-1)^2 & 4 & & (-1)^{J+1} \\ & & \ddots & \vdots \\ (-1)^J & (-1)^{J+1} & \cdots & 4 \end{bmatrix}$$

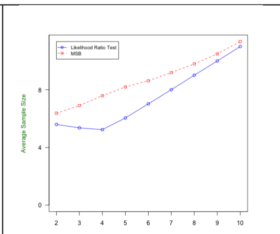
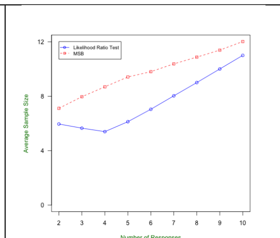
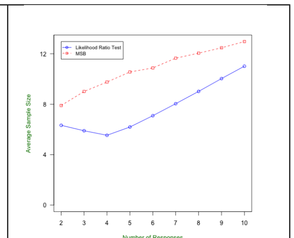
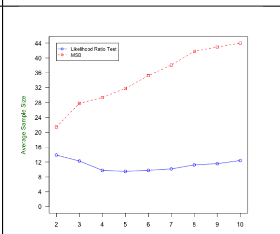
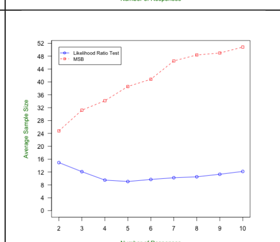
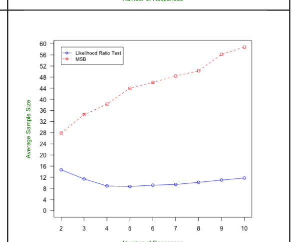
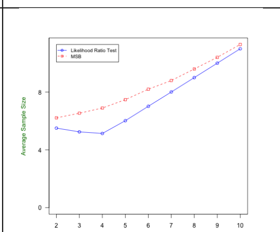
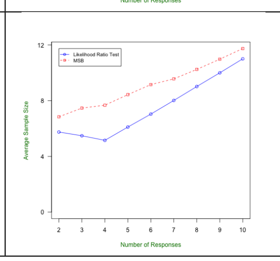
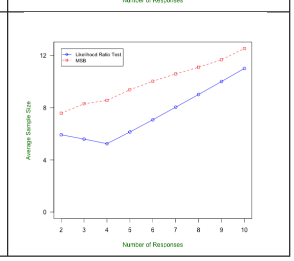
Table 2 and Table 3 present the average sample sizes required by our proposed likelihood ratio test and MSB. Table 2 shows the average samples and achieved errors for Likelihood Ratio Test and MSB when $J = 5$. Notice that we have many "-" in Table 3 since, for instance, if the factor is unimportant, we cannot commit Type II error. Table 3 compares the performance of likelihood ratio test with MSB under different value of J, the number of responses. We vary J from 2 to 10. Let Σ be the same and define β as

$$\beta_1' = (0,0,...,0,1) \qquad \beta_2' = (0,0,...,0,2) \qquad \beta_3' = (0,0,...,0,5)$$

Table 2: Average sample size for a single-factor screening.

β	Covariance	Likelihood Ratio Test			MSB		
		Avg Sample Size	Type I Error	Type II Error	Avg Sample Size	Type I Error	Type II Error
$(0,0,0,0,1)$	Σ_1	6.154	0.1%	-	9.513	0.0%	-
	Σ_2	6.163	0.2%	-	9.456	0.0%	-
	Σ_3	6.100	0.9%	-	9.549	0.0%	-
$(1,1,1,1,1)$	Σ_1	6.536	2.2%	-	11.895	0.0%	-
	Σ_2	6.495	3.4%	-	11.598	0.1%	-
	Σ_3	6.418	3.3%	-	11.954	0.0%	-
$(0,0,0,0,3)$	Σ_1	9.818	-	-	36.606	-	-
	Σ_2	9.121	-	-	38.307	-	-
	Σ_3	9.214	-	-	39.753	-	-
$(3,3,3,3,3)$	Σ_1	6.350	-	-	22.766	-	-
	Σ_2	7.008	-	-	27.678	-	-
	Σ_3	6.174	-	-	22.097	-	-
$(0,0,0,0,5)$	Σ_1	6.141	-	0.0%	8.535	-	0.1%
	Σ_2	6.063	-	0.2%	8.278	-	0.0%
	Σ_3	6.073	-	0.3%	8.405	-	0.0%
$(5,5,5,5,5)$	Σ_1	6.000	-	0.0%	6.192	-	0.0%
	Σ_2	6.001	-	0.0%	6.296	-	0.0%
	Σ_3	6.000	-	0.0%	6.201	-	0.0%

Table 3: Average sample size for single-factor screening.

	Σ_1	Σ_2	Σ_3
β_1'			
β_2'			
β_3'			

We can see the proposed likelihood ratio test are more efficient than MSB. The advantage of the likelihood ratio test over MSB is even more obvious when factors are within indifferent zone. In the five response case, likelihood ratio test saves 30% simulation effort on average when compared with MSB. When the factor is important but not critical, likelihood ratio test needs only approximately $1/4$ of the computation effort of MSB. Table 3 indicates that in some case, the gap between likelihood ratio test and MSB is closing as the number of responses is increasing. But likelihood ratio test still performs universally better than MSB in terms of computational effort. Moreover, for MSB, the computational effort highly depends on the coefficient of the factors and the average sample size varies from 10 to 40 when $j = 10$. As for likelihood ratio test, the average sample size is between 10 to 12 when $j = 10$.

4.2 Multiple-Factors Multiple-Responses Study

This experiment is to compare SUMIP and SORTIP with HIP. Resolution III fractional factorial design is used in all three procedures. Consider a factor screening problem with five responses. Table 4 lists all parameters for this experiment.

Table 4: Simulation experiment parameter of 2.

Parameter	Value	Meaning
I	50, 100, and 150	Number of Factors
J	4	Number of Responses
Δ_0	2	Threshold of Important Factors
Δ_1	4	Threshold of Critial Factors
α	0.05	Upperbound of FWERI
γ	0.1	Upperbound of FWERII
β	Specified below	Responses
Σ	Specified below	Covariance matrix
n_0	5	Initial sample size

We consider two scenarios for β_{ij}. Scenario one has 5% critical factors, 5% important but not critical factors, and 90% unimportant factors. Scenario two has 10% critical factors, 10% important but not critical factors, and 80% unimportant factors. Critical factors' coefficients are randomly generated from uniform distribution on $(\Delta_1, 6)$, unimportant factors' coefficients are generated from uniform distribution $(0, \Delta_0)$, the important factors are generated from uniform distribution (Δ_0, Δ_1).

For the i^{th} factor, its covariance is one of the three cases,

$$\Sigma_i = \begin{bmatrix} 4 & 1 & 1 & 1 \\ 1 & 4 & 1 & 1 \\ 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 4 \end{bmatrix} + m_i \begin{bmatrix} \beta_{i1} & 0 & 0 & 0 \\ 0 & \beta_{i2} & 0 & 0 \\ 0 & 0 & \beta_{i3} & 0 \\ 0 & 0 & 0 & \beta_{i4} \end{bmatrix}$$

where $m_1 = 0$, $m_2 = 0.5$ and $m_2 = 2$.

We randomly generate two scenarios of β and Table 5 and Table 6 present the average sample sizes of 1,000 independent experiments required by SUMIP, SORTIP, and HIP. The sample size of each experiment is the product of two parts: (1) the number of factors in the Resolution III fractional factorial design, (2) the number of replications to reach a conclusion. For instance, if we have 50 factors and 10 replications to reach the conclusion, since the number of the design points in Resolution III fractional factorial design is 64, the sample size would be 64 times 10. In all cases in these two scenarios, both SUMIP and SORTIP dominate HIP in terms of average sample size. SUMIP and SORTIP use approximately 50% to 60% sample size required in HIP in most cases without the loss in power. Although SUMIP perform little worse than SORTIP, its stopping criteria is much simpler than both SORTIP's and HIP's.

Table 5: Average sample size for HIP, SUMIP, and SORTIP in Scenario One (5% factors are critical).

		HIP			SUMIP			SORTIP		
I	Σ	Avg Samples	FWERI	FWERII	Avg Samples	FWERI	FWERII	Avg Samples	FWERI	FWERII
50	Σ_1	1038.208	0.0%	0.0%	520.512	0.0%	0.5%	474.816	0.0%	1.0%
	Σ_2	1089.216	0.0%	0.6%	481.984	0.0%	0.7%	470.848	0.0%	0.7%
	Σ_3	1179.264	0.0%	0.2%	535.808	0.0%	0.3%	535.920	0.0%	0.3%
100	Σ_1	2583.040	0.0%	0.0%	1229.560	0.0%	0.6%	1113.984	0.0%	0.6%
	Σ_2	2597.120	0.0%	0.0%	1136.896	0.0%	0.6%	1078.144	0.0%	0.6%
	Σ_3	2923.392	0.0%	0.0%	1298.048	0.0%	1.3%	1189.130	0.0%	1.3%
150	Σ_1	10965.504	0.0%	0.2%	5142.528	0.0%	0.2%	4904.960	0.0%	0.2%
	Σ_2	11048.960	0.0%	0.1%	5204.480	0.0%	0.5%	4874.240	0.0%	0.5%
	Σ_3	12914.416	0.0%	0.1%	6337.536	0.0%	0.2%	5766.144	0.0%	0.2%

Table 6: Average sample size for HIP, SUMIP, and SORTIP in Scenario Two (10% factors are critical).

		HIP			SUMIP			SORTIP		
I	Σ	Avg Samples	FWERI	FWERII	Avg Samples	FWERI	FWERII	Avg Samples	FWERI	FWERII
50	Σ_1	696.448	0.0%	0.0%	401.920	0.0%	0.6%	390.208	0.0%	0.6%
	Σ_2	828.032	0.0%	0.1%	470.976	0.0%	0.7%	458.944	0.0%	0.7%
	Σ_3	1049.088	0.0%	0.0%	636.608	0.0%	0.1%	603.84	0.0%	0.2%
100	Σ_1	1752.576	0.0%	0.0%	993.024	0.0%	0.1%	923.136	0.0%	0.3%
	Σ_2	2418.176	0.0%	0.0%	1300.350	0.0%	0.6%	1243.008	0.0%	0.7%
	Σ_3	3097.600	0.0%	0.3%	1704.704	0.0%	0.9%	1623.424	0.0%	0.9%
150	Σ_1	7746.54	0.0%	0.0%	4401.664	0.0%	0.1%	4039.680	0.0%	0.2%
	Σ_2	11716.096	0.0%	0.0%	6277.12	0.0%	0.0%	6024.704	0.0%	0.0%
	Σ_3	15515.136	0.0%	0.1%	7777.792	0.0%	0.2%	7432.193	0.0%	0.2%

5 CONCLUSION

SUMIP and SORTIP are sequential multiple responses factor screening procedures that provide strong controls on FWERI and FWERII simultaneously. With the option of using fractional factorial designs, SUMIP and SORTIP can handle large-scale problems efficiently. Numerical evaluation indicates the performances of SUMIP and SORTIP are robust and efficient across different simulation configurations.

Our future research will concentrate on developing sequential bifurcation or grouping procedure that controls the FWERI and FWERII. In addition, since previous research of simulation factor screening focused on controlling the error rate with economical designs, a related research topic would be the optimal computational budget allocations, which is how to allocate design budgets in order to minimize the error rates and maximize the power of the tests.

REFERENCES

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

Elster, C., and A. Neumaier. 1995, September. "Screening by conference designs". *Biometrika* 82 (3): 589–602.

Kleijnen, J. P., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. 2005. "State-of-the-art review: a user's guide to the brave new world of designing simulation experiments". *INFORMS Journal on Computing* 17 (3): 263–289.

Lee, L. H., E. P. Chew, and S. Teng. 2007. "Finding the Pareto set for multi-objective simulation models by minimization of expected opportunity cost". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 513–521. Piscataway, New Jersey.

Lee, M.-L. T. 2004. *Analysis of Microarray Gene Expression Data*. Springer.

Logan, B. R. "A cone order monotone test for the one-sided multivariate testing problem". *Statistics & Probability Letters* 63 (3): 315–323.

Myers, R. H., C. M. Anderson-Cook, and D. C. Montgomery. 2009. *Response surface methodology: process and product optimization using designed experiments*. Hoboken, NJ: Wiley.

Perlman, M. D., and L. Wu. 2004. "A Note on One-Sided Tests with Multiple Endpoints". *Biometrics* 60 (1): 276–280.

Perlman, M. D., and L. Wu. 2006. "Some Improved Tests for Multivariate One-Sided Hypotheses". *Metrika* 64 (1): 23–39.

Reitmeir, P., and G. Wassmer. 1996. "One-sided multiple endpoint testing in two-sample comparisons". *Communications in Statistics - Simulation and Computation* 25 (1): 99–117.

Saltelli, A., S. Tarantola, and F. Campolongo. 2000. "Sensitivity Analysis as an Ingredient of Modeling". *Statistical Science* 15 (4): 377–395.

Shen, H., and H. Wan. 2009. "Controlled sequential factorial design for simulation factor screening". *European Journal of Operational Research* 198 (2): 511–519.

Shen, H., H. Wan, and S. M. Sanchez. 2010. "A hybrid method for simulation factor screening". *Naval Research Logistics* 57 (1): 45–57.

Shi, W., J. P. Kleijnen, and Z. Liu. 2014. "Factor screening for simulation with multiple responses: Sequential bifurcation". *European Journal of Operational Research* 237 (1): 136–147.

Trocine, L., and L. C. Malone. 2000. "Statistical tools for simulation design and analysis I: finding important independent variables through screening designs: a comparison of methods". In *Proceedings of the 2000 Winter Simulation Conference*, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 749–754. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Wald, A., and J. Wolfowitz. 1948. "Optimum Character of the Sequential Probability Ratio Test". *The Annals of Mathematical Statistics* 19 (3): 326–339.

Wan, H., and B. E. Ankenman. 2007. "Two-stage controlled fractional factorial screening for simulation experiments". *Journal of Quality Technology* 39 (2): 126–139.

Wan, H., B. E. Ankenman, and B. L. Nelson. 2006. "Controlled sequential bifurcation: A new factor-screening method for discrete-event simulation". *Operations Research* 54 (4): 743–755.

AUTHOR BIOGRAPHIES

WENYU WANG is a Ph.D. student in the School of Industrial Engineering at Purdue University. His research interests include simulation optimization and stochastic modeling. His email address is wang1708@purdue.edu.

HONG WAN is an Associate Professor in the School of Industrial Engineering at Purdue University. Her research interests include design and analysis of simulation experiments, simulation optimization, applied statistics, quality management, and healthcare engineering. She has taught a variety of courses and is a member of INFORMS and ASA. Her email address is hwan@purdue.edu and her web page is http://web.ics.purdue.edu/hwan.