Two-stage Sensitivity-based Group Screening in Computer Experiments

Hyejung Moon, Thomas Santner, Angela Dean
(hjmoon@stat.osu.edu) (tjs@stat.osu.edu) (amd@stat.osu.edu)
Department of Statistics, The Ohio State University
1958 Neil Avenue, Columbus, OH 43210

Abstract

Sophisticated computer codes that implement mathematical models of physical processes can take hours to produce a single response. Screening to determine the most active inputs is critical for understanding the input-output relationship. This paper presents new methodology based on two-stage group screening. In stage 1, groups of inputs are screened out and, at stage 2, individual active inputs are sought. Inputs are evaluated through their total effect sensitivity indices (TSIs) which are compared with a benchmark null TSI distribution. Examples show that, in comparison with with one-stage procedures, the proposed method provides accurate screening while reducing computational effort.

KEY WORDS: Active factor; Experimental design; Gaussian process; Latin hypercube design; Low-impact input; Total sensitivity index

1 INTRODUCTION

1.1 Background

A computer model is a numerical implementation of a mathematical description of an input-output relationship. Such models are prevalent, in a wide range of applications; for example, in engineering (Fang, Li, and Sudjianto (2005)), biomechanics (Ong, Lehman, Notz, Santner,

and Bartel (2006)), physical sciences (Higdon, Kennedy, Cavendish, Cafeo, and Ryne (2004)), and life sciences (Fogelson, Kuharsky, and Yu (2003)).

Over the past 20 years, the use of computer codes as experimental tools has become increasingly sophisticated, and can allow the user to vary environmental and calibration inputs in addition to inputs that describe different "treatments" (see, for example, Ong, Santner, and Bartel (2008)). Thus, computer simulators based on sophisticated finite-element or computational fluid dynamic numerical methods can require hours or even days for a single run. There is a need, therefore, for efficient methods to detect *active* or *influential* inputs that have *major* impacts on an input-output system. Once identified, researchers can restrict attention to varying these inputs, while setting other inputs to nominal values, thus reducing the complexity of the emulator to the maximum extent.

The literature contains several proposals for screening inputs in computer experiments based on the assumption that the deterministic output is modeled as a realization of a random function. An approach that decomposes the output from the computer simulator into main effects and interaction effects, has been applied by many authors. For example, Sacks, Welch, Mitchell, and Wynn (1989), Welch et al. (1992), Oakley and O'Hagan (2004), Schonlau and Welch (2006), and Morris, Moore, and McKay (2008) use the random function model in various ways to estimate these effects. An alternative approach by Linkletter, Bingham, Hengartner, Higdon, and Ye (2006) (hereafter called LBHHY) selects active inputs based on draws from the posterior distribution of the correlation parameters of a stationary Gaussian process model with Gaussian correlation function.

This paper extends the use of the *group screening* methodology from the physical experiments setting to that of computer simulators, and proposes a computationally efficient method of identifying active inputs. Group screening methodology was first described by Dorfman (1943) for blood screening and later adapted to *physical experiments* by Watson (1961) for identifying active factors (variables) in the presence of many potentially influential

factors. For more recent work in this area, see Morris and Mitchell (1983), Lewis and Dean (2001), Vine, Lewis, and Dean (2005), and reviews by Kleijnen (1987) and Morris (2006)).

In two-stage group screening in physical experiments, factors are commonly investigated at only two settings, where the "high" ("low") level corresponds to the setting expected to result in the higher (lower) response. Factors are placed into groups and the high (low) level of the "grouped factor" occurs when all individual factors within that group are at their high (low) levels. A first stage experiment is run on the small number of grouped factors and the analysis screens out groups of factors that have little effect on the response; all factors within such a group are declared *inactive*. A second-stage experiment is run on the individual factors within the active groups, while the inactive factors are held at nominal levels. The analysis makes a final selection of active individual factors from the set of potentially active factors. An example of such a screening experiment run in the automobile industry is described in detail by Vine, Lewis, Dean, and Brunson (2008).

In the *computer experiment* setting, the response surface is often highly complex and, for a Gaussian random function predictor to be a well-fitting emulator for the computer simulator, design points are required to be well-spaced across the experimental region, thus utilizing a large number of levels for each input variable. Due to the added complexity, it is not obvious how to group the variables in such a way as to achieve an efficient screening procedure. Thus, in our proposed screening procedure, GSinCE (Group Screening in Computer Experiments), described in Section 1.2, we incorporate a grouping phase in which we use data from the computer simulator to suggest a good grouping of input variables and also to fit a Bayesian Gaussian Process model for generating data for screening. For clarity of exposition, we divide the description of each stage of GSinCE into sampling, grouping (for Stage 1 only), and analysis phases. Sensitivity analyses are performed to select the active inputs. Due to the deterministic nature of the response, the second-stage analysis is able to use the initial set of data (taken prior to grouping) from the simulator as well as the second stage data.

GSinCE enables inputs having small effects (not solely zero effects) to be eliminated.

All of our codes for the screening procedures were written specifically for this paper in the programming language MATLAB; they call the GPM/SA code (Gaussian Process Models for Simulation Analysis) of Gattiker (2005) for estimation of parameters in the Gaussian Process model and to compute sensitivity indices.

1.2 Overview of the Proposed Procedure

Initialization Given that n runs of the computer code are to be made in Stage 1, we first generate a matrix $\boldsymbol{X}^*$ with n rows and $(n-1)$ columns satisfying certain desirable properties, described in Section 2. The choice of n is based on the anticipated maximum number of active inputs (see Section 3.1). The columns of $\boldsymbol{X}^*$ are generated to be centered and orthogonal and, consequently, $\boldsymbol{X}^*$ cannot have more than $(n-1)$ columns. The columns of $\boldsymbol{X}^*$ are used for two purposes at stage 1; first to provide a design for the $f < n-1$ input variables of interest and, second, to provide a design for the m grouped input variables and up to $n-m-1$ benchmark inputs against which the activity of the group input variables will be measured. If n is very large, there may be more benchmark columns available than are needed by our procedure; in this case, we recommend that $\boldsymbol{X}^*$ be generated with $min(n-1, f+50)$ columns instead of the $(n-1)$ columns assumed throughout the paper.

Stage 1 In the sampling phase, a set of columns from $\boldsymbol{X}^*$ is selected to produce a design matrix $\boldsymbol{X}^{(1)}$. The computer simulator is run at the design points (rows) in $\boldsymbol{X}^{(1)}$ and a Gaussian process (GP) model is fitted to the output as an emulator for the code. In the grouping phase, the output is used to place the inputs into disjoint groups. All inputs in the same group are set equal to the *same* level, defined by a design matrix $\boldsymbol{G}$ (Section 3.2) and the fitted GP model is used to predict the output at the design points in $\boldsymbol{G}$. The analysis phase (Section 3.3) uses *total effect sensitivity indices* to determine which groups of inputs are inactive and which potentially contain active inputs. To judge whether a group is active

or non-active, an additional "low-impact" input is created to use as a "benchmark" (c.f. LBHHY and Wu, Boos, and Stefanski (2007)).

Stage 2 The inputs in the groups selected as active in Stage 1 are investigated individually in Stage 2. In the Stage 2 sampling phase (Section 4.1), a new design matrix $\boldsymbol{X}^{(2)}$ is selected in such a way that the design points in the combined $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$ retain, as closely as possible, the desirable properties identified in Section 2. The computer simulator is run at the design points in $\boldsymbol{X}^{(2)}$. The Stage 2 analysis phase uses the outputs from both stages in a second sensitivity analysis to make the final selection of active inputs (Section 4.2).

Detailed descriptions of the Initialization, Stage 1, and Stage 2 are given in Sections 2, 3 and 4, respectively. Among the issues addressed by these sections are: (i) the desirable properties for design matrices, (ii) grouping strategies, (iii) the creation of low-impact inputs, and (iv) the determination of active inputs via sensitivity analysis. Section 5 demonstrates the methodology via simulated examples and proposes the optimal magnitude of the low-impact input by studying certain operating characteristics of the procedure. Section 6 presents two examples that demonstrate the performance of GSinCE and compare it with a one-stage sensitivity analysis and the one-stage screening procedure of LBHHY. It is shown that the GSinCE procedure is computationally efficient and remarkably stable in its selection of active inputs. Lastly, Section 7 states our conclusions, including circumstances in which screening can be problematic for any procedure, as well as possible extensions of the GSinCE procedure. We note that LBHHY, GSinCE and the one-stage procdure are all based on a reasonable fit of a Gaussian Process model and estimates of sensitivity indices which assume that the likelihood is Gaussian. If it were not possible to fit such a model, then one possible approach to estimating sensitivity indices would be a sampling based method such as those described by Morris, Moore, and McKay (2006).

2 GSinCE INITIALIZATION STAGE

Suppose that the range of the j^{th} input is $[a_j, b_j]$, with a_j and b_j known constants, $j = 1, \ldots, f$, and that the domain for the vector of the f inputs is the entire hyper-rectangle $\prod_{j=1}^{f}[a_j, b_j]$. We obtain the design from the scaled input space $[0,1]^f$ and let $x_{ij} \in [0,1]$ be the value of the j^{th} *scaled input* in the i^{th} run of the design, for $i = 1, \ldots, n$ and $j = 1, \ldots, f$. The computer simulator is run to obtain the output using the *unscaled input*.

In the Initialization Stage, we construct a preliminary design matrix $\boldsymbol{X}^*$; we denote the j^{th} column of $\boldsymbol{X}^*$ by $\boldsymbol{\xi}_j = (\xi_{1j}, \ldots, \xi_{nj})^\top$, $j = 1, \ldots, n-1$, where $\top$ denotes transpose, and the i^{th} row of $\boldsymbol{X}^*$ as $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{i(n-1)})$, $i = 1, \ldots, n$. The design matrices for the Stage 1 sampling and grouping phases will be drawn from this matrix, as will those for the low-impact inputs.

We make three requirements for $\boldsymbol{X}^*$, as follows: (P.1) the columns of $\boldsymbol{X}^*$ must be uncorrelated to allow independent assessment of the effects of the different inputs; (P.2) the minimum and maximum values in each column must be 0 and 1, respectively, to prevent input values with larger ranges from having larger impacts on the response, artificially induced by the design; (P.3) the design $\boldsymbol{X}^*$ should be "space-filling" whose purpose is to insure that all regions of the input space are explored (c.f. Sacks et al. (1989), Santner, Williams, and Notz (2003), ch. 5). There are several ways to define "space-filling" designs; for example, maximize the minimum Euclidean interpoint distance measured in $(n-1)$-dimensional space, or minimize the average reciprocal distance between design points in any user-selected collection of subspaces of the $(n-1)$-dimensional space (see Welch (1985)). Here, we select a criterion that maximizes the minimum interpoint distance within all 2-dimensional subspaces of the input space. Moon, Dean, and Santner (2011) show that not only does the the construction of such designs save considerable computing time, but the designs tend to perform well also under a maximin criterion in the $(n-1)$-dimensional space. The following algorithm generates an $\boldsymbol{X}^*$ that satisfies (P.1), (P.2), and approximately satisfies (P.3).

Step 1 Generate an $n \times (n-1)$ Latin hypercube design matrix $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{n-1})$, where $\boldsymbol{\lambda}_i$ is a random permutation of $\{1, 2, \ldots, n\}$, (see McKay, Beckman, and Conover (1979)).

Step 2 Center each column of $\boldsymbol{\Lambda}$: $\boldsymbol{v}_h = \boldsymbol{\lambda}_h - (\boldsymbol{\lambda}_h^\top \mathbf{1}/n)\mathbf{1}$ for $h = 1, \ldots, n-1$, where $\mathbf{1}$ is a vector of n unit elements.

Step 3 Apply the Gram-Schmidt algorithm to form orthogonal columns $\boldsymbol{u}_h = (u_{1h}, \ldots, u_{nh})^\top$,

$$\boldsymbol{u}_h \;=\; \begin{cases} \boldsymbol{v}_1, & h = 1; \\ \boldsymbol{v}_h - \sum_{i=1}^{h-1} \frac{\boldsymbol{u}_i^\top \boldsymbol{v}_h}{||\boldsymbol{u}_i||^2} \boldsymbol{u}_i, & h = 2, \ldots, n-1. \end{cases}$$

If any $\boldsymbol{u}_h$ is zero, the original $\boldsymbol{\Lambda}$ is not full-rank and is re-generated.

Step 4 Scale the values of $\boldsymbol{u}_h$ to [0,1] to give $\boldsymbol{\xi}_h$ $(h = 1, \ldots, n-1)$. Set $\boldsymbol{X} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n-1})$.

Selection of the design $\boldsymbol{X}^* = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n-1})$ which maximizes the minimum inter-point distance over all projections into 2-dimensional space can be achieved (approximately) by repeating Steps 1–4 many times and finding the best among the candidate designs generated. Alternatively, a genetic exchange algorithm could be used, as discussed by Moon et al. (2011).

3 GSinCE PROCEDURE STAGE 1

3.1 Stage 1 Sampling Phase

The GSinCE procedure is to be used in a screening situation where it is reasonable to assume that only a small fraction (say 25% or less) of the inputs are active. Loeppky, Sacks, and Welch (2009) justified "10 × number of inputs" as a rule of thumb for the number of runs in an effective initial computer experiment. Using this base value, we suggest 5 runs for each active input in each stage, so with a conservative assumption of a maximum of 40% active inputs, we take $n = 5 \times (f \times 0.4) = 2f$ runs in Stage 1. To simplify notation, we write the Stage 1 design matrix, $\boldsymbol{X}^{(1)}$, as the first f columns of $\boldsymbol{X}^*$; thus $\boldsymbol{X}^{(1)} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_f)$. A Bayesian GP model (see Higdon et al. (2004))

$$Y(\boldsymbol{x}) = Z(\boldsymbol{x}) + \epsilon(\boldsymbol{x}) \tag{1}$$

is used to fit the computer simulator data $\boldsymbol{y}(\boldsymbol{X}^{(1)})$. Here $Z(\cdot)$ is taken to be a stationary Gaussian process with zero mean, and covariance function

$$Cov(Z(\boldsymbol{x}), Z(\tilde{\boldsymbol{x}})) = \frac{1}{\lambda_Z} R(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{1}{\lambda_Z} \prod_{j=1}^{f} \rho_j^{4(x_j - \tilde{x}_j)^2} \ , \tag{2}$$

where $\boldsymbol{x} = (x_1, \ldots, x_f)$ and $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_f)$ are two design points. The term $\epsilon(\boldsymbol{x})$ in (1) represents numerical or other small scale noise and is modeled by a white noise process that is independent of $Z(\cdot)$ and has mean 0 and (small) prior variance $1/\lambda_\epsilon$. The output $\boldsymbol{y}(\boldsymbol{X}^{(1)})$ is centered to have sample mean 0 and unit variance to conform to the prior specification when fitting this Bayesian model. The model can be fitted using the GPM/SA software. The posterior distributions of the model parameters will be used to predict output for the group variables in the grouping phase in Section 3.2.

3.2 Stage 1 Grouping Phase

Since the objective of the screening procedure is to screen out those inputs that have little effect on the outputs, an optimal grouping scheme would place these inputs into the same group. Similarly, grouping of inputs that have similar effects (e.g. linear or quadratic, increasing or decreasing) on the response is helpful to avoid masking the effects of some inputs by other inputs. Information from subject experts is extremely valuable in grouping together the inputs that have similar behavior. Additionally, exploratory data analysis of the Stage 1 inputs and outputs can be used to suggest groupings.

In the context of our simulation study in Section 5, it was not possible to use expert opinion for the grouping. Consequently, we developed an automated grouping procedure based on the Pearson correlation coefficients $r(\boldsymbol{\xi}_j, \boldsymbol{y}(\boldsymbol{X}^{(1)}))$, $(j = 1, \ldots, n)$ which measure the strength of the linear relationship between the inputs and the output (see the on-line supplement for details). We used the automated procedure also in our example in Section 6.1

and found that it works well even though the responses are non-linear in the inputs. Through a 60-input example in Section 6.2, we discuss a few differences obtained with user-determined groups and the automated procedure. A further comparison of expert and automatic groupings in an experiment run at Los Alamos National Laboratory and involving 61 inputs and 500 runs is described in Section 4.2 of Moon (2010).

After the f individual inputs have been divided into m, say, groups, a design matrix $\boldsymbol{G} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m)$ is formed from m randomly selected columns from $\boldsymbol{X}^*$ for the group variables. From $\boldsymbol{G}$, design matrix $\boldsymbol{X}^P = (\boldsymbol{\xi}_1^P, \ldots, \boldsymbol{\xi}_f^P)$ is constructed in terms of the f individual inputs, where *all* the inputs in group k are set to the levels defined by $\boldsymbol{g}_k$, $k = 1, \ldots, m$. For example, if inputs 1, 5, 6 are assigned to group 1, then $\boldsymbol{\xi}_1^P = \boldsymbol{\xi}_5^P = \boldsymbol{\xi}_6^P = \boldsymbol{g}_1$. The design matrix $\boldsymbol{X}^P$ is used to predict the output based on the fitted GP model (Section 3.1). The resulting values, $\widehat{\boldsymbol{y}}(\boldsymbol{G})$, are used in Section 3.3.2 to select the active groups. The training data, $\boldsymbol{y}(\boldsymbol{X}^{(1)})$, will be used again in Section 4.2 to help select active individual inputs within the active groups.

3.3 Stage 1 Analysis Phase

3.3.1 Sensitivity Indices

In Sections 3.3.2 and 4.2, total sensitivity indices (TSIs) are used to detect active effects. This subsection reviews the definition of sensitivity indices when the input region is $[0,1]^f$. Sobol´ (1993) showed that the function $y(\boldsymbol{x})$ can be decomposed as

$$y(\boldsymbol{x}) = y_0 + \sum_{j=1}^{f} y_j(x_j) + \sum_{1 \le j < h \le f} y_{jh}(x_j, x_h) + \ldots + y_{1,2,\ldots,f}(x_1, \ldots, x_f) \tag{3}$$

where the terms are recursively defined by

$$\begin{aligned} y_0 &= \int_{[0,1]^f} y(x_1, \ldots, x_f) dx_1 \ldots dx_f \\ y_j(x_j) &= \left[\int_{[0,1]^{f-1}} y(x_1, \ldots, x_f) d\boldsymbol{x}_{-j} \right] - y_0 \\ y_{jh}(x_j, x_h) &= \left[\int_{[0,1]^{f-2}} y(x_1, \ldots, x_f) d\boldsymbol{x}_{-jh} \right] - y_j(x_j) - y_h(x_h) - y_0 \end{aligned}$$

and so on. Here $d\boldsymbol{x}_{-j}$ denotes integration over all inputs *except* x_j, and $d\boldsymbol{x}_{-jh}$ denotes integration over all inputs *except* x_j and x_h. The individual components of the decomposition are centered and orthogonal; that is, they satisfy $\int_0^1 y_{j_1,\ldots,j_s}(x_{j_1},\ldots,x_{j_s})dx_{j_k} = 0$, for any $1 \leq k \leq s$, and $\int_{[0,1]^f} y_{j_1,\ldots,j_s}(x_{j_1},\ldots,x_{j_s})y_{h_1,\ldots,h_t}(x_{h_1},\ldots,x_{h_t})dx_1\ldots dx_f = 0$, for any $(j_1,\ldots,j_s) \neq (h_1,\ldots,h_t)$. Variance-based indices with respect to the uniform distribution over $[0,1]^f$ for $\boldsymbol{X}$ are obtained by squaring both sides of (3) and integrating (Sobol´ (1993)). This leads to the variance decomposition:

$$V = \sum_{j=1}^{f} V_j + \sum_{1\leq j<h\leq f} V_{jh} + \cdots + V_{1,2,\ldots,f} \tag{4}$$

$$\text{where } V = \left[\int_{[0,1]^f} y^2(x_1,\ldots,x_f)dx_1\ldots dx_f\right] - y_0^2,$$

$$V_j = \int_0^1 y_j^2(x_j)dx_j, \quad \text{and} \quad V_{jh} = \int_0^1\int_0^1 y_{jh}^2(x_j,x_h)dx_jdx_h,$$

and additional terms are defined similarly. Sensitivity indices are obtained by dividing each component in (4) by the total variance V. The *main effect sensitivity index* of the j^{th} input, is defined to be $S_j = V_j/V$, and the *two-factor sensitivity index* of the j^{th} and h^{th} inputs is defined to be $S_{jh} = V_{jh}/V$. Higher-order sensitivity indices are defined similarly. The TSI of the j^{th} input (Homma and Saltelli (1996)) is the sum of all sensitivity indices involving the j^{th} input; that is,

$$T_j = S_j + \sum_{h\neq j} S_{jh} + \ldots + S_{1,2,\ldots,f}. \tag{5}$$

In this paper, sensitivity indices are computed using the Bayesian method of Oakley and O'Hagan (2004) as implemented in GPM/SA; the sensitivity index is estimated by the mean of the posterior distribution which is obtained from the posterior draws of the GP model parameters via Markov chain Monte Carlo (MCMC).

3.3.2 A Benchmark Null Distribution for Sensitivity Indices

LBHHY used a reference distribution for variable selection obtained by augmenting the experimental inputs with an input known to be inert. This approach is attractive since it removes the need for subjective assessment about which inputs have "large" indicators of activity. Augmentation with an input which has zero effect on the output can lead to the selection of inputs with very small effects as being active. In practice, we wish to eliminate low-impact inputs as well as totally inert ones. LBHHY addressed this issue in their discussion and suggested that the response could be spiked with some very small effect (see also, Wu et al. (2007)). We use the latter approach in this paper and modify the output by adding an input having a small, user-determined, effect. Any group of inputs whose TSI is smaller than that of the added low-impact input is treated as non-active.

LBHHY selected random columns from the input space for the inert input. However, since randomly generated columns can be correlated with the columns of the design matrix for the inputs of interest, this paper instead draws multiple, uncorrelated columns from the preliminary design matrix $\boldsymbol{X}^*$ for the low-impact inputs. There are $(n-1)-m$ columns of $\boldsymbol{X}^*$ that are both uncorrelated with the m columns of $\boldsymbol{G}$ and with each other. Thus, by augmenting $\boldsymbol{G}$ with these $n-m-1$ columns in turn, we construct $n-m-1$ augmented group design matrices. Denote the w^{th} such design matrix by

$$\boldsymbol{G}(w) = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m, \boldsymbol{g}_{m+1}^{(w)})$$

where $\boldsymbol{g}_i$ is the column of $\boldsymbol{G}$ for the i^{th} group variable, $1 \leq i \leq m$, and $\boldsymbol{g}_{m+1}^{(w)}$ is the column selected from $\boldsymbol{X}^*$ for the w^{th} low-impact input, $1 \leq w \leq n-m-1$. Each $\boldsymbol{G}(w)$ satisfies the properties (P.1) and (P.2) and approximately (P.3) of Section 2.

We set the magnitude of the low-impact input to be a fraction τ, $0 < \tau < 1$, of the range of the output; choices for τ that optimize the performance of the GSinCE procedure are discussed in Section 5. Let

$$\beta = \left(\max_{1 \leq t \leq n} \hat{y}_t - \min_{1 \leq t \leq n} \hat{y}_t \right) \tau \tag{6}$$

define the magnitude of the effect of the low-impact input, where $\hat{y}_t$ is the t^{th} value in the predicted output $\widehat{\boldsymbol{y}}(\boldsymbol{G})$ defined in Section 3.2. For the design $\boldsymbol{G}(w)$, a perturbation of the predicted data $\widehat{\boldsymbol{y}}(\boldsymbol{G})$ is computed using $\boldsymbol{g}_{m+1}^{(w)}$, to give

$$\hat{\boldsymbol{y}}^{(w)} = \widehat{\boldsymbol{y}}(\boldsymbol{G}) + \beta\, \boldsymbol{g}_{m+1}^{(w)}. \tag{7}$$

The perturbed output $\hat{\boldsymbol{y}}^{(w)}$ is used to compute the TSIs in (5) corresponding to the m group inputs and the w^{th} low-impact input; these quantities are denoted by $T_1^{(w)}, \ldots, T_m^{(w)}, T_{m+1}^{(w)}$, for $1 \leq w \leq n - m - 1$.

The GSinCE procedure selects the i^{th} group variable as being active if its TSI is larger than that of the low-impact input. This decision is made based on the pairs of TSIs $(T_{m+1}^{(w)}, T_i^{(w)})$, $1 \leq w \leq n - m - 1$, using the sign test (see Conover (1999)). The $n - m - 1$ pairs $(T_{m+1}^{(w)}, T_i^{(w)})$ are mutually independent because each pair of TSIs is estimated by the posterior mean of independent MCMC draws given $\hat{\boldsymbol{y}}^{(w)}$. We assume the pairs are internally consistent, in that if $P[T_{m+1}^{(w)} < T_i^{(w)}] \leq P[T_{m+1}^{(w)} > T_i^{(w)}]$ for one w then it is true for all w. Then the hypotheses for the ith TSI, $1 \leq i \leq m$, are formulated as

$$\begin{aligned} H_{0i} &: P[T_{m+1}^{(w)} < T_i^{(w)}] \leq P[T_{m+1}^{(w)} > T_i^{(w)}], \quad \text{for all } w \\ H_{1i} &: P[T_{m+1}^{(w)} < T_i^{(w)}] > P[T_{m+1}^{(w)} > T_i^{(w)}], \quad \text{for all } w \end{aligned} \tag{8}$$

and tested at significance level α/m to account for the multiple groups, yielding a "family-wise" significance level α.

4 GSinCE PROCEDURE STAGE 2

4.1 Stage 2 Sampling Phase

Suppose that there are p inputs in total in the groups identified as active at Stage 1; each of these p inputs is potentially active. Using a sample size justification similar to that in Section 3.1, we take $n_1 = 5(f \times 0.4) = 2f$ runs for Stage 1 and $n_2 = 5p$ runs for Stage 2, using conservative assumptions of the proportion of active inputs. Rearrange the

columns of $\boldsymbol{X}^*$ to firstly list the p columns associated with potentially active inputs (A), then those for the $(f-p)$ non-active inputs (N), followed by the low-impact benchmark (B) columns, to obtain the matrix $(\boldsymbol{X}_A^{(1)}, \boldsymbol{X}_N^{(1)}, \boldsymbol{X}_B^{(1)})$. The maximum number of uncorrelated low-impact benchmark columns that can be included in the design matrix $\boldsymbol{X}^c$ defined below is $\delta_B = \min(n_1 - f - 1, n_2 - p - 1)$.

Next, we construct an $n_2 \times p$ stage 2 design matrix $\boldsymbol{X}_A^{(2)}$ for the inputs from the active groups, an $n_2 \times (f-p)$ matrix $\boldsymbol{X}_N^{(2)}$ for the inputs from the non-active groups, and an $n_2 \times \delta_B$ matrix $\boldsymbol{X}_B^{(2)}$ for the low-impact inputs. The values in the jth column of $\boldsymbol{X}_N^{(2)}$ are all set equal to the median value of the jth column of $\boldsymbol{X}_N^{(1)}$. The $n_2 \times (p + \delta_B)$ matrix $(\boldsymbol{X}_A^{(2)}, \boldsymbol{X}_B^{(2)})$ is constructed using Steps 1–4 of the design algorithm in Section 2 so that its columns satisfy (P.1) and (P.2); Step 5 is implemented so that the *combined* $(n_1 + n_2) \times (p + \delta_B)$ matrix

$$\boldsymbol{X}^c = \begin{pmatrix} \boldsymbol{X}_A^{(1)} & \boldsymbol{X}_B^{(1)} \\ \boldsymbol{X}_A^{(2)} & \boldsymbol{X}_B^{(2)} \end{pmatrix} = (\boldsymbol{X}_A^c, \boldsymbol{X}_B^c)$$

is (approximately) maximin. The columns of $\boldsymbol{X}^c$ will satisfy (P.2) but need not be uncorrelated (P.1). However, in the examples that we have investigated, their correlations are very small. The computer code is now run at each set of input values defined by the rows of $\boldsymbol{X}^{(2)} = (\boldsymbol{X}_A^{(2)}, \boldsymbol{X}_N^{(2)})$ to obtain the output $\boldsymbol{y}(\boldsymbol{X}^{(2)})$.

4.2 Stage 2 Analysis Phase

The n_1 output values $\boldsymbol{y}(\boldsymbol{X}^{(1)})$ from Stage 1 are used together with the n_2 output values $\boldsymbol{y}(\boldsymbol{X}^{(2)})$ from Stage 2; let $\boldsymbol{y}^c = (\boldsymbol{y}(\boldsymbol{X}^{(1)})^\top, \boldsymbol{y}(\boldsymbol{X}^{(2)})^\top)^\top = (y_1^c, \ldots, y_{n_1+n_2}^c)^\top$ denote the combined data. We construct δ_B augmented design matrices by appending, one-by-one, the (low-impact) columns $\boldsymbol{\xi}_{p+1}^{c(w)}$ $(w = 1, \ldots, \delta_B)$ in $\boldsymbol{X}_B^c$ to the combined design matrix for the potentially active factors, $\boldsymbol{X}_A^c$. We represent the w^{th} such design by

$$\boldsymbol{X}^c(w) = (\boldsymbol{X}_A^c,\ \boldsymbol{\xi}_{p+1}^{c(w)})$$

for $1 \leq w \leq \delta_B$. The effect of the w^{th} low-impact input $\boldsymbol{\xi}_{p+1}^{c(w)}$ is based on the combined

output,

$$\beta^c = \left(\max_{1 \le t \le (n_1+n_2)} y_t^c \; - \min_{1 \le t \le (n_1+n_2)} y_t^c \right) \; \tau \tag{9}$$

using the same value of τ selected at Stage 1. The perturbed output for the Stage 2 analysis using $\boldsymbol{\xi}_{p+1}^{c(w)}$ is defined as

$$\boldsymbol{y}^{c(w)} = \boldsymbol{y}^c + \beta^c \; \boldsymbol{\xi}_{p+1}^{c(w)}. \tag{10}$$

The output $\boldsymbol{y}^{c(w)}$ is used to compute the TSIs $T_1^{(w)}, \ldots, T_p^{(w)}, T_{p+1}^{(w)}$ corresponding to the p individual inputs and the w^{th} low-impact input. We obtain δ_B pairs of TSIs to test the activity level for any given individual input among the p potentially active inputs. Following the procedure used in Section 3.3, the hypotheses to test that the j^{th} input is active are formulated as

$$\begin{aligned} H_{0j} &: P[T_{p+1}^{(w)} < T_j^{(w)}] \le P[T_{p+1}^{(w)} > T_j^{(w)}], \quad \text{for all } w \\ H_{1j} &: P[T_{p+1}^{(w)} < T_j^{(w)}] > P[T_{p+1}^{(w)} > T_j^{(w)}], \quad \text{for all } w \end{aligned} \tag{11}$$

and tested at significance level α/p, for $j = 1, \ldots, p$, for a selected familywise significance level α.

5 SIMULATION STUDIES TO SET τ

We now determine a setting of τ in (6) and (9) to control the operating characteristics of the GSinCE procedure, using a stochastic test bed of second-order polynomials:

$$y(z_1, \ldots, z_f) = \sum_{j=1}^{f} \gamma_j z_j + \sum_{j=1}^{f} \sum_{h=j}^{f} \; \gamma_{jh} z_j z_h, \tag{12}$$

where $z_j \in [0, 1]$, $j = 1, \ldots, f$. The study reported here uses $f = 20$ factors, $n_1 = 2f$ runs in stage 1, and approximately 25% active factors. Moon (2010) extends the study to $(f, n_1) \in \{(10, 20), (30, 60), (20, 80)\}$ and 35%, 20% and fewer active factors. The hypothesis tests (8) and (11) are performed with a familywise significance level $\alpha = 0.2$ at each stage. The large value $\alpha = 0.2$ is selected since it is deemed less important to select inactive inputs than to (incorrectly) screen out active ones.

Let L, Q, and I denote the set of inputs involved in active linear, quadratic, and interaction effects, respectively. The values of the regression coefficients γ_j, γ_{jj}, and γ_{jh} are assigned under the principles of effect sparsity, hierarchy, and heredity, described in Chipman (2006). For $j = 1, \ldots, f$, the linear coefficient γ_j is drawn from the following distribution,

$$\gamma_j \sim \begin{cases} N(\mu_A, \sigma_A^2), & \text{with prob } q_L, \\ N(\mu_N, \sigma_N^2), & \text{with prob } 1 - q_L, \end{cases} \tag{13}$$

where $q_L = P[j \in L]$ is the probability that the linear effect of input j is active and $\mu_A > 0$, $\mu_A > \mu_N$, $\sigma_A > \sigma_N$. Let $q_{Q|A}$ and $q_{Q|N}$ be the conditional probabilities that the quadratic effect of input j is active given that the linear effect is active or non-active, i.e., $q_{Q|A} = P[j \in Q | j \in L]$ and $q_{Q|N} = P[j \in Q | j \notin L]$. We draw γ_{jj} from the distribution,

$$\gamma_{jj} \sim \begin{cases} N(\mu_A, \sigma_A^2) & \text{with prob } \frac{q_Q}{2}, \\ N(-\mu_A, \sigma_A^2) & \text{with prob } \frac{q_Q}{2}, \\ N(\mu_N, \sigma_N^2) & \text{with prob } 1 - q_Q. \end{cases} \quad \text{where } \; q_Q = \begin{cases} q_{Q|A}, & \text{if } j \in L, \\ q_{Q|N}, & \text{if } j \notin L. \end{cases} \tag{14}$$

Similarly, we let $q_{\times|AA}$, $q_{\times|AN}$, $q_{\times|NA}$, $q_{\times|NN}$ be the conditional probabilities that the interaction between inputs j and h is active given that the linear effects of each of these inputs is active or non-active. Then γ_{jh} is drawn from the distribution,

$$\gamma_{jh} \sim \begin{cases} N(\mu_A, \sigma_A^2) & \text{with prob } \frac{q_\times}{2}, \\ N(-\mu_A, \sigma_A^2) & \text{with prob } \frac{q_\times}{2}, \\ N(\mu_N, \sigma_N^2) & \text{with prob } 1 - q_\times. \end{cases} \quad \text{where } \; q_\times = \begin{cases} q_{\times|AA}, & \text{if } j \in L, h \in L, \\ q_{\times|AN}, & \text{if } j \in L, h \notin L, \\ q_{\times|NA}, & \text{if } j \notin L, h \in L, \\ q_{\times|NN}, & \text{if } j \notin L, h \notin L. \end{cases} \tag{15}$$

It is straightforward to show that the expected proportion of active inputs is

$$q_L + q_{Q|N}(1 - q_L) + (f - 1)\left(q_{\times|AN} q_L + q_{\times|NN}(1 - q_L)\right)(1 - q_{Q|N})(1 - q_L). \tag{16}$$

Table 1 lists four sets of marginal probabilities, labelled $P_1 - P_4$, that were selected to generate second-order polynomials with the expected proportion of active inputs (16) approximately equal to 0.25. P_1 produces the largest number of active linear effects and the

fewest active quadratic and interaction effects. P_3 produces more active quadratic and interaction effects whose linear effects are also active. P_2 and P_4 produce more active quadratic and interaction effects whose corresponding linear effect is not active, but fewer active linear effects. P_4 is more extreme in this regard. Thus, P_2, P_3 and P_4 can create more complicated test functions than P_1, while having a similar expected proportion of active inputs. Table 1 also lists three sets of coefficient distributions labelled $C_1 - C_3$. Normal distributions were selected for the active effects and the distribution for the non-active effects is fixed as $N(0, 2^2)$. The coefficients drawn for the active effects under C_1 are the easiest to differentiate from those of the non-active effects, and those under C_2 are the hardest.

Table 1: Marginal probabilities and coefficient distributions for the simulation study

Choice of Marginal Probabilities							Choice of Coefficient Distributions				
	q_L	$q_{Q\|A}$	$q_{Q\|N}$	$q_{\times\|AA}$	$q_{\times\|AN}$	$q_{\times\|NN}$		μ_A	σ_A^2	μ_N	σ_N^2
P_1	0.15	0.10	0.005	0.10	0.010	0.005	C_1	40	10^2	0	2^2
P_2	0.10	0.10	0.010	0.10	0.014	0.008	C_2	20	5^2	0	2^2
P_3	0.15	0.90	0.005	0.90	0.010	0.005	C_3	30	7.5^2	0	2^2
P_4	0.05	0.10	0.050	0.10	0.010	0.009					

Combinations of P_i and C_j were used to generate a wide variety of functions. Of the twelve possible combinations, the six listed in Table 2 proved to be the most challenging for screening.

Table 2: Six combinations used to recommend τ

Combination	1	2	3	4	5	6
Marginal Probabilities	P_2	P_2	P_3	P_3	P_4	P_4
Coefficient Distributions	C_2	C_3	C_2	C_3	C_2	C_3

In a screening problem, there are two kinds of errors; we can falsely select non-active inputs, or falsely not select active inputs. The false discovery rate (*FDR*, see Benjamini and Hochberg (1995)) and false non-discovery rate (*FNDR*) are defined as:

- $\text{FDR} = \dfrac{\text{number of non-active inputs that are claimed to be active}}{\text{number of inputs claimed to be active}}$
- $\text{FNDR} = \dfrac{\text{number of active inputs that are claimed to be non-active}}{\text{number of inputs claimed to be non-active}}$

Two standard positive performance measures are *specificity* and *sensitivity* (see, for example, Altman and Bland (1994)) which are defined to be:

- specificity $= \dfrac{\text{number of true non-active inputs that are claimed to be non-active}}{\text{number of true non-active inputs}}$
- sensitivity $= \dfrac{\text{number of true active inputs that are claimed to be active}}{\text{number of true active inputs}}$

When the denominator is 0, then FDR or FNDR is defined to be 0 and specificity or sensitivity is defined to be 1. Our objective is to select a value of τ for the low-impact inputs as used in (6) and (9) which leads to low FDR and FNDR and high specificity and sensitivity.

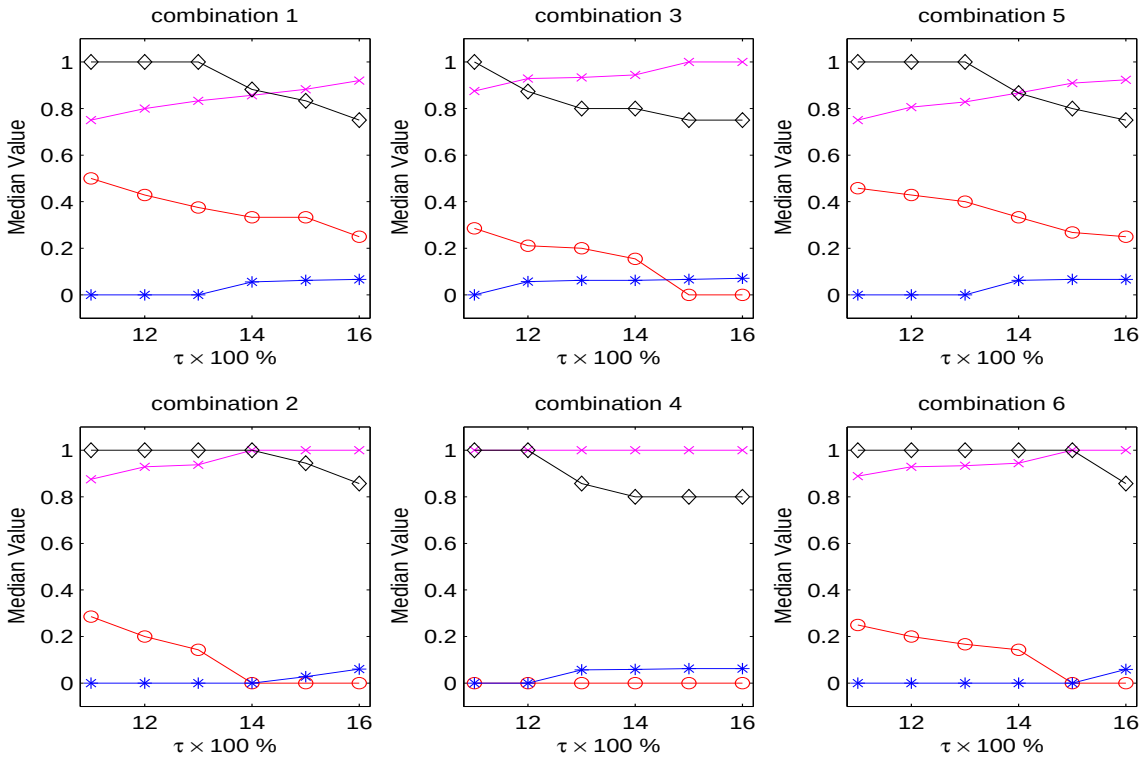

Figure 1: Median Values of FDR (line with circle), FNDR (line with asterisk), specificity (line with cross), sensitivity (line with diamond) over 200 Test Functions versus $\tau \times 100\%$.

The GSinCE procedure was applied to each of 200 randomly generated second-order polynomials in each of the six combinations of Table 2 using $\tau \in \{0.11, 0.12, 0.13, 0.14, 0.15, 0.16\}$. Stage 1 grouping was performed by the automated grouping procedure with a maximum

group size $M = 5$. In Figure 1, we plot the median values of FDR, FNDR, specificity, sensitivity for combinations 1–6 of Table 2, when 25% of the total inputs are active on average, and $f = 20$, $\alpha = 0.2$. For these combinations, the values $\tau \in [0.13, 0.15]$ seem to give a reasonable compromise between low median FDR and FNDR and high median specificity and sensitivity. Moon (2010) shows that this range of τ values remains reasonable for 25% active inputs for the $f = 10$ and $f = 30$ cases, and also for an expected 35% active inputs and $f = 10$. For a smaller percentage of active inputs, Moon (2010) recommends increasing τ to approximately 0.2 at Stage 2 to reduce the otherwise large number of false positives. (Some details are given in the on-line supplement).

6 EXAMPLES

The proposed GSinCE procedure is now illustrated for two examples and compared with (i) a one-stage method that uses TSI test (11) with $p = f$, $\tau = 0.14$, and the same number of runs as GSinCE, and (ii) the LBHHY procedures using $\tau = 0$ and $\tau = 0.14$. Following the example in LBHHY, the initial design for the LBHHY procedures is selected using the maximin LHD criterion, and 100 randomly selected columns are added to form a benchmark distribution; the procedure judges whether an input is active by the magnitude of the posterior draws of the parameters of the correlation function $R(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ in (2), and the 10th percentile of the benchmark distribution is selected as the cut-off for the selection decision.

6.1 Borehole Model

Worley (1987) used the function

$$y(z_1, \ldots, z_8) = \frac{2\pi z_3 (z_4 - z_6)}{\ln(z_2/z_1)\left[1 + \frac{2z_7 z_3}{\ln(z_2/z_1) z_1^2 z_8} + \frac{z_3}{z_5}\right]}$$

to describe the rate of flow of water (in m^3/yr) through a borehole that is drilled from the ground surface through two aquifers, where $z_1 \in [0.05, 0.15]$ is the radius of the borehole (m); $z_2 \in [100, 50000]$ is the radius of influence (m); $z_3 \in [63070, 115600]$ is the transmissivity

of upper aquifer (m^2/yr); $z_4 \in [990, 1110]$ is the potentiometric head of the upper aquifer (m); $z_5 \in [63.1, 116]$ is the transmissivity of the lower aquifer (m^2/yr); $z_6 \in [700, 820]$ is the potentiometric head of the lower aquifer (m); $z_7 \in [1120, 1680]$ is the length of the borehole (m); $z_8 \in [9855, 12045]$ is the hydraulic conductivity of the borehole (m/yr). In their illustration of a Bayesian variable selection technique for this model, Joseph, Hung, and Sudjianto (2008) identified z_1 as the only important variable among the 8 inputs based on 27-run experimental design.

To illustrate GSinCE, we add twelve inert inputs $z_9, \ldots, z_{20}$ which serve only to add "noise" to the variable screening process. Then, in our analysis $y(\cdot)$, is considered as a function of $f = 20$ inputs. We use $n_1 = 2f = 40$ runs at Stage 1. Using the Stage 1 data, the automated grouping procedure (see the on-line supplementary material), with maximum group size $M = 5$, places the inputs into 7 groups as shown in Table 3; for example, g_1 consists of inputs z_6 and z_7. The GSinCE analysis phase with $\tau = 0.14$ and hypothesis test (8) with $\alpha = 0.2$, selects groups g_1, g_6, and g_7. The $p = 6$ inputs in these groups, $(z_1, z_4, z_6, z_7, z_8, z_{19})$, proceed to Stage 2. Setting $n_2 = 5p = 30$, the 30×6 design $\boldsymbol{X}^{(2)}$ is obtained as in Section 4.1. Following the computation of the 30 additional code runs to obtain $\boldsymbol{y}(\boldsymbol{X}^{(2)})$, the inputs z_1, z_4, z_6, z_7 are selected as active in Stage 2, three more than selected by Joseph et al. (2008) (who did not consider the additional 12 inert inputs).

Table 3: Results of applying GSinCE to the borehole model

r_j^* range	Individual Inputs	Group	Stage 1 Selection	Stage 2 Selection
-0.25 to -0.21	z_7, z_6	g_1	$\surd$	z_6, z_7
-0.07 to -0.03	$z_{17}, z_3, z_{10}, z_{14}$	g_2		
-0.02 to -0.00	$z_{13}, z_9, z_{12}, z_{11}$	g_3		
0.02	$z_{16}, z_{18}, z_2, z_{15}$	g_4		
0.03 to 0.04	z_{20}, z_5	g_5		
0.07 to 0.20	z_{19}, z_8, z_4	g_6	$\surd$	z_4
1.49	z_1	g_7	$\surd$	z_1

Table 4 compares the GSinCE procedure with a 70-run one-stage sensitivity analysis

and the two LBHHY procedures. The three procedures that use $\tau = 0.14$ to construct the benchmark select the same set of inputs to be active. However, the GSinCE procedure is considerably faster than the three one-stage procedures, (the computation times measure the execution for the MATLAB codes used to implement the four procedures on the same 64 bit Linux machine with 8 cores, 32 GB of RAM, and 2.66 GHz). When used with an inert benchmark ($\tau = 0$), LHBBY additionally selects input z_8, and also declares the inert input z_{19} to be active. This shows the importance of using a non-inert benchmark to screen out inputs that are essentially noise. As a check on the selection of active inputs, we computed TSIs using training data collected from a fine input grid. The TSIs are 0.866, 0.053, 0.053, 0.053, 0.012, for inputs z_1, z_4, z_7, z_6, z_8 respectively, compared with the largest TSI for the inert inputs of 0.00065.

Table 4: Results of applying four Procedures to the borehole model (70 runs)

Procedure	τ	Selected Inputs	Computation Times (1,000 seconds)
GSinCE	0.14	z_1, z_4, z_6, z_7	3.0
One-stage	0.14	z_1, z_4, z_6, z_7	7.1
LBHHY	0.14	z_1, z_4, z_6, z_7	4.6
LBHHY	0	$z_1, z_4, z_6, z_7, z_8, \boldsymbol{z_{19}}$	4.5

6.2 A 60-input example

We created a complex function with $f = 60$ inputs, where z_{32}–z_{60} are known to be inert. The output function $y(\cdot)$ is given in detail in the on-line supplement. A computation of TSIs using training data over a fine grid resulted in TSIs under 0.0127 for the added inert variables. The largest TSIs for the remaining inputs were 0.214, 0.154, 0.132, 0.124, 0.101, 0.099, for $z_1, z_{26}, z_{19}, z_{20}, z_{16}, z_{27}$, respectively. The next largest TSIs were for inputs z_{11}, z_{15} with values 0.058, 0.044, respectively, followed by $z_9, z_{17}, z_{25}, z_{21}$ with TSI values in the range 0.025–0.028.

For Stage 1 of GSinCE, with $n_1 = 2f = 120$, we compared the results of the automated grouping with $M = 5$ and two groupings that might have been chosen by examining the

correlations of the input values with the output (which we call "expert groupings"). The automated grouping results in 18 groups, some of which are listed in Table 5. The expert grouping combined groups 17 and 18 together. Expert grouping 2 additionally moved input z_1, which has the largest correlation with the output, into a new group 18. The groups selected by GSinCE using test (8) with $\tau = 0.14$ are are shown in the penultimate row of Table 5. These pass to Stage 2 where $n_2 = 2p$ additional code runs are evaluated.

Table 5: Results of different groupings for the 60-run example

	Automatic Grouping	Expert Grouping 1	Expert Grouping 2
Grouping	g1 = (19, 26) g2 =(15, 6, 7) ⋮ g16=(17, 21, 9, 4) g17 =(11) g18=(27, 16, 20, 1)	g1 = (19, 26) g2 =(15, 6, 7) ⋮ g16=(17, 21, 9, 4) g17=(11, 27,16, 20,1)	g1 = (19, 26) g2 =(15, 6, 7) ⋮ g16=(17, 21, 9, 4) g17=(11, 27, 16, 20) g18=(1)
Selected Groups	g1, g2, g16, g18	g1, g2, g16, g17	g1, g2, g16, g17, g18
# inputs selected, p	13	14	14

The top section of Table 6 shows the results of Stage 2 of GSinCE using both the expert grouping and the automated grouping. Also shown are the results from the one-stage sensitivity analysis with the same total number of runs and the LBHHY procedures with $\tau = 0.14$ and $\tau = 0$. It is clear that LBHHY with $\tau = 0$ selects many inputs with very small sensitivity indices, including some of the inert inputs $z_{32} - z_{60}$. All procedures with $\tau = 0.14$ are similar and select 6–8 inputs having the largest TSIs.

If the computer code is slow to run, it is desirable to use as few code runs as possible when screening. Consequently, we investigated the effect of reducing both n_1 and n_2 on the success of each procedure in determining the active inputs. Table 6 shows the results of reducing n_2 from the recommended $5p$ to $4p, 3p, 2p$, in GSinCE, while retaining $n_1 = 2f = 120$. With the expert grouping, the results are remarkably stable, where the six inputs with the largest TSIs are selected, except for the one case in the automated grouping where the sixth is omitted. The one-stage and LBHHY procedures with $\tau = 0.14$ and the same number of total runs as

Table 6: Results of varying n_2, when $n_1 = 120$ for expert and automated groupings

Total Runs	Method	Expert Grouping 1		Total Runs	Automated Grouping	
		Inputs Selected	n		Inputs Selected	n
$(5p)$ 190	GSinCE	1, 26, 19, 20, 16, 27	6	$(5p)$ 185	1, 26, 19, 20, 16, 27	6
	One-stage	1, 26, 19, 20, 16, 27, 11, 15	8		1, 26, 19, 20, 16, 27	6
	LBHHY	1, 26, 19, 20, 16, 27	6		1, 26, 19, 20, 16, 27	6
	LBHHY ($\tau = 0$)	1, 4, 6-9, 11, 14-17, 19-22, 25-29, 43, 56, 59	23		1, 4, 6-9, 11, 15-17, 19-22, 25-28	18
$(4p)$ 176	GSinCE	1, 26, 19, 20, 16, 27	6	$(4p)$ 172	1, 26, 19, 20, 16, 27	6
	One-stage	1, 26, 19, 20, 16, 27, 11	7		1, 26, 19, 20, 16, 27, 11	7
	LBHHY	1, 26, 19, 20, 16, 27, 11	7		1, 26, 19, 20, 16, 27, 28	7
	LBHHY ($\tau = 0$)	1, 4, 6-9, 11, 15-22, 25-28, 44, 51	21		1, 4, 6-11, 15-22, 25-30, 45, 47, 54	25
$(3p)$ 162	GSinCE	1, 26, 19, 20, 16, 27	6	$(3p)$ 159	1, 26, 19, 20, 16	5
	One-stage	1, 26, 19, 20, 16, 27, 11, 15, 9	9		1, 26, 19, 20, 16, 27, 11	7
	LBHHY	1, 26, 19, 20, 16, 27, 11	7		1, 26, 19, 20, 16, 27, 11	7
	LBHHY ($\tau = 0$)	1, 4, 6-9, 11, 15-22, 25-28, 32, 34	21		1, 4, 6-9, 11, 15-17, 19-23, 25-28, 33	20
$(2p)$ 148	GSinCE	1, 26, 19, 20, 16, 27	6	$(2p)$ 146	1, 26, 19, 20	4
	One-stage	1, 26, 19, 20, 16, 27	6		1, 26, 19, 20, 16, 27, 11	7
	LBHHY	1, 26, 19, 20, 16, 27	6		1, 26, 19, 20, 16, 27	6
	LBHHY ($\tau = 0$)	1, 3, 4, 6, 7, 9, 11, 14-21, 25-28	19		1, 4, 6-9, 11, 15-22, 25-28, 49, 57	21

GSinCE are slightly more variable, although LBHHY always selcts 6 or 7 inptus with the largest TSIs.

Next, we investigated the results of reducing n_1 from $5 \times 0.4f = 120$ to $5 \times 0.3f = 90$ to $5 \times 0.23f = 70$ in combination with $n_2 = 5p$ and $n_2 = 2p$. As n_1 decreased, the number of low-impact benchmark inputs at Stage 1 consequently decreased from 59 to 29 to 9. The results are shown in Table 7 for expert grouping 1. When $n_1 = 90$, GSinCE selected groups containing $p = 15$ inputs for at Stage 1, so that $n_2 = 5p = 75$ or $n_2 = 2p = 30$ runs were added at Stage 2. For $n_1 = 70$, GSinCE selected groups containing $p = 8$ inputs at Stage 1, so $n_2 = 5p = 40$ or $n_2 = 2p = 16$ runs were added at Stage 2. For $n_1 = 90$, GSinCE selects the seven inputs with the largest TSIs, but becomes less stable when n_1 drops to 70.

For the one-stage sensitivity analysis, the number of low-impact columns was set to 29 or 9 to match Stage 1 of the GSinCE procedure. This analysis selected 6–9 inputs with

the largest TSIs. Following LBHHY, we added 100 columns for the low-impact benchmark inputs for their procedures. With 165 runs, LBHHY selected z_{28} (TSI=0.0165) instead of z_{16} (TSI=0.1018), and otherwise selected the 5–8 inputs with the largest TSIs.

In summary, GSinCE seems to be the most stable of the procedures examined. For a small total number of runs, we recommend that n_1 be retained at $2f$ when 15% to 25% of inputs are expected to be active and, if necessary, n_2 can be reduced to as small as $2p$.

Table 7: Results of reducing n_1 to 70 and 90 for expert grouping 1

Total Runs	Method	Selected	
		Inputs	n
165($5p$) $n_1 = 90$	GSinCE	1, 11, 16, 19, 20, 26, 27	7
	One-stage	1, 16, 19, 20, 26, 27	6
	LBHHY	1, 19, 20,26, 27, 28	6
	LBHHY ($\tau = 0$)	1-4, 6-9, 11, 14-17, 19-22, 25-29, 37, 39, 52	25
120($2p$) $n_1 = 90$	GSinCE	1, 11, 16, 19, 20, 26, 27	7
	One-stage	1, 11, 16, 19, 20, 26, 27	7
	LBHHY	1, 11, 15, 16, 19, 20, 26, 27	8
	LBHHY ($\tau = 0$)	1, 4, 6-9, 11, 15-22, 25-28, 50	20
110($5p$) $n_1 = 70$	GSinCE	1, 16, 19, 26, 27	5
	One-stage	1, 11, 16, 19, 20, 26, 27	7
	LBHHY	1, 11, 16, 19, 20, 26, 27	7
	LBHHY ($\tau = 0$)	1, 4, 6-9, 11, 15-17, 19-22, 25-27, 57	18
96($2p$) $n_1 = 70$	GSinCE	1, 9, 11, 16, 19, 20, 26, 27	8
	One-stage	1, 9, 11, 15, 16, 19, 20, 26, 27	9
	LBHHY	1, 16, 19, 20, 26	5
	LBHHY ($\tau = 0$)	1, 4-7, 9, 11, 15-17, 19-21, 25-28, 40	18

7 SUMMARY AND DISCUSSION

This paper presents a two-stage statistical methodology, GSinCE, for identifying the active inputs in computer models with high-dimensional inputs. The use of low-impact benchmark inputs allows GSinCE to have high specificity and low FDR. GSinCE has an advantage over other procedures by being computationally efficient, since the number of inputs is reduced through grouping at Stage 1 and only potentially active inputs are examined

at Stage 2. We have found that when the proportion of active inputs is in the range 15% to 25%, accurate and stable results are obtained by using $n_1 = 2f$ runs at Stage 1 and $n_2 = 5p$ at Stage 2. However, if the budget is tight so that $n_1 + n_2$ is bounded, it appears to be important to retain $n_1 = 2f$ runs at stage 1 with consequently fewer runs at Stage 2. This is likely to be because GSinCE uses data from both stages in the Stage 2 sensitivity analysis.

We generate our designs using Gram-Schmidt orthogonalization to achieve orthogonal columns, and select from among these under a maximin criterion over all 2-dimensional projections. Other criteria can be used, and one possibility is to use projections into k-dimesional space, where k is an upper bound for the anticipated number of active inputs.

When the first stage data $y(\boldsymbol{X})$ have previously been collected according to some design $\boldsymbol{X}$ before screening became the goal, our procedure would use these data at the grouping phase to fit the Gaussian Process model. Other aspects of GSinCE would remain the same, except that it may be difficult to satisfy (P.2) approximately for $\boldsymbol{X}^c$ unless the original design were orthogonal.

There are a number of circumstances in which any type of screening is problematic and, in particular, in which modifications of GSinCE can improve its operating characteristics. First, when the $\boldsymbol{x} \to y(\boldsymbol{x})$ relationship is non-linear, the sensitivity of GSinCE can be improved by increasing the number of runs at Stage 1. Second, if $y(\cdot)$ has very few or even no active inputs, GSinCE with $\tau = 0.14$ at both stages can falsely identify inputs having small effects as being active, but this can be avoided by increasing τ at Stage 2.

The most problematic case is when $y(\cdot)$ has numerous non-active inputs each having small effects of the *same* sign. Here GSinCE can perform poorly in terms of specificity due to the amalgamation of effects of the factors within the groups. Even in this case, however, GSinCE can still identify active inputs whose effects are sufficiently large to be separated from the non-active inputs.

ACKNOWLEDGMENTS

This research was sponsored, in part, by the National Science Foundation under Agreement DMS-0806134 (The Ohio State University). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Altman, D. G. and Bland, J. M. (1994), "Diagnostic tests. 1: Sensitivity and specificity," *British Medical Journal*, 308, 1552.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300.

Chipman, H. (2006), "Prior Distributions for Bayesian Analysis of Screening Experiments," in *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, eds. Dean, A. M. and Lewis, S. M., New York: Springer Verlag, pp. 235–267.

Conover, W. J. (1999), *Practical Nonparametric Statistics*, John Wiley & Sons.

Dorfman, R. (1943), "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, 14, 436–440.

Fang, K. T., Li, R., and Sudjianto, A. (2005), *Design and Modeling for Computer Experiments*, London: Chapman and Hall.

Fogelson, A., Kuharsky, A., and Yu, H. (2003), "Computational Modeling of Blood Clotting: Coagulation and Three-dimensional Platelet Aggregation," in *Polymer and Cell Dynamics: Multicsale Modeling and Numerical Simulations*, eds. Alt, W., Chaplain, M., Griebel, M., and Lenz, J., Basel: Birkhauser-Verlag, pp. 145–154.

Gattiker, J. (2005), "Using the Gaussian Process Model for Simulation Analysis (GPM/SA) Code," Tech. Rep. LA-UR-05-5215, Los Alamos National Laboratory.

Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., and Ryne, R. (2004), "Combining field data and computer simulations for calibration and prediction," *SIAM Journal of Scientific Computing*, 26, 448–466.

Homma, T. and Saltelli, A. (1996), "Importance measures in global sensitivity analysis of model output," *Reliability Engineering and System Safety*, 52, 1–17.

Joseph, V. R., Hung, Y., and Sudjianto, A. (2008), "Blind Kriging: A New Method for Developing Metamodels," *ASME Journal of Mechanical Design*, 130, 031102–1–8.

Kleijnen, J. P. C. (1987), "Review of Random and Group-screening Designs," *Communications in Statistics: Theory and Methods*, 16, 2885–2900.

Lewis, S. M. and Dean, A. M. (2001), "Detection of Interactions in Experiments on Large Numbers of Factors," *Journal of the Royal Statistical Society, Ser. B, (with discussion)*, 63, 633–672.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. Q. (2006), "Variable Selection for Gaussian Process Models in Computer Experiments," *Technometrics*, 48, 478–490.

Loeppky, J. L., Sacks, J., and Welch, W. (2009), "Choosing the Sample Size of a Computer Experiment: A Practical Guide," *Technometrics*, 51, 366–376.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979), "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, 21, 239–245.

Moon, H. (2010), "Design and Analysis of Computer Experiments for Screening Input Variables," Ph.D. thesis, The Ohio State University.

Moon, H., Dean, A. M., and Santner, T. J. (2011), "Algorithms for Generating Maximin Latin Hypercube and Orthogonal Designs," *Journal of Statistical Theory and Practice*, 5, 81–98.

Morris, M. D. (2006), "An Overview of Group Factor Screening," in *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, eds. Dean, A. M. and Lewis, S. M., New York: Springer Verlag, pp. 191–206.

Morris, M. D. and Mitchell, T. J. (1983), "Two-level Multifactor Designs for Detecting the Presence of Interactions," *Technometrics*, 25, 345–355.

Morris, M. D., Moore, L. M., and McKay, M. D. (2006), "Sampling Plans Based on Balanced Incomplete Block Designs for Evaluating the Importance of Computer Model Inputs," *Journal of Statistical Planning and Inference*, 136, 3203–3220.

— (2008), "Using Orthogonal Arrays in the Sensitivity Analysis of Computer Models," *Technometrics*, 50, 205–215.

Oakley, J. E. and O'Hagan, A. (2004), "Probabilistic Sensitivity Analysis of Complex Models: A Bayesian Approach," *Journal of the Royal Statistical Society, Ser. B*, 66, 751–769.

Ong, K., Lehman, J., Notz, W. I., Santner, T. J., and Bartel, D. L. (2006), "Acetabular Cup Geometry and Bone-Implant Interference have More Influence on Initial Periprosthetic Joint Space than Joint Loading and Surgical Cup Insertion," *Journal of Biomechanical Engineering*, 148, 169–175.

Ong, K., Santner, T. J., and Bartel, D. L. (2008), "Robust Design for Acetabular Cup Stability Accounting for Patient and Surgical Variability," *Journal of Biomechanical Engineering*, 130, 031001–11.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and analysis of computer experiments," *Statistical Science*, 4, 409–423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer Verlag.

Schonlau, M. and Welch, W. J. (2006), "Screening the Input Variables to a Computer Model Via Analysis of Variance and Visualization," in *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, eds. Dean, A. M. and Lewis, S. M., New York: Springer Verlag, pp. 308–327.

Sobol´, I. M. (1993), "Sensitivity analysis for non-linear mathematical models," *Mathematical Modeling and Computational Experiment*, 1, 407–414.

Vine, A. E., Lewis, S. M., and Dean, A. M. (2005), "Two-stage Group Screening in the Presence of Noise Factors and Unequal Probabilities of Active Effects," *Statistica Sinica*, 15, 871–888.

Vine, A. E., Lewis, S. M., Dean, A. M., and Brunson, D. (2008), "A Critical Assessment of Two-Stage Group Screening Through Industrial Experimentation," *Technometrics*, 50, 15–25.

Watson, G. S. (1961), "A Study of the Group Screening Method (Com: V5 P397-398; V7 P444-446)," *Technometrics*, 3, 371–388.

Welch, W. J. (1985), "ACED: Algorithms for the Construction of Experimental Designs," *The American Statistician*, 39, 146–146.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992), "Screening, predicting, and computer experiments," *Technometrics*, 34, 15–25.

Worley, B. A. (1987), "Deterministic Uncertainty Analysis," ORNL-6428, available from National Technical Information Service, 5285 Port Royal Road, Spring VA 22161.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007), "Controlling Variable Selection by the Addition of Pseudovariables," *Journal of the American Statistical Association*, 102, 235–243.