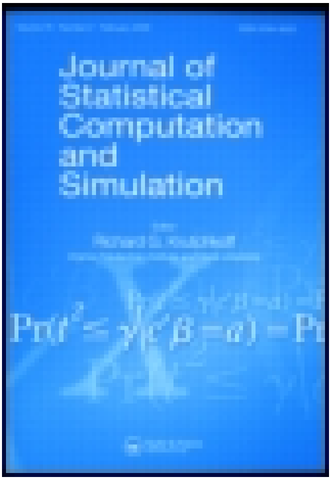

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Sensitivity measures,anova-like Techniques and the use of bootstrap

G. E. B. Archer [a], A. Saltelli [a] & I. M. Sobol [b]

[a] Joint Research Centre of the European Commission, Ispra, VA, 21020, Italy

[b] National Centre for Mathematical Modelling of the Russian Academy of Science, 4A Miusskaya Square, 125047, Moscow(CIS)

Published online: 20 Mar 2007.

To cite this article: G. E. B. Archer, A. Saltelli & I. M. Sobol (1997) Sensitivity measures,anova-like Techniques and the use of bootstrap, Journal of Statistical Computation and Simulation, 58:2, 99-120, DOI: 10.1080/00949659708811825

To link to this article: http://dx.doi.org/10.1080/00949659708811825

J. Statist. Comput. Simul., 1997, Vol. 58, pp. 99–120
Reprints available directly from the publisher
Photocopying permitted by license only

SENSITIVITY MEASURES, ANOVA-LIKE TECHNIQUES AND THE USE OF BOOTSTRAP

G. E. B. ARCHER[a], A. SALTELLI[a] and I. M. SOBOL[b]

[a]*Joint Research Centre of the European Commission, 21020 Ispra (VA), Italy;*
[b]*National Centre for Mathematical Modelling of the Russian Academy of Science, 4A Miusskaya Square, 125047 Moscow (CIS)*

(Received 3 September 1996; In final form 21 January 1997)

Sobol′ sensitivity indices, used in variance based global sensitivity analysis of model output, are compared with the Analysis of Variance in classical factorial design. Monte Carlo computation of Sobol′ indices is described briefly, and a bootstrap approach is presented, which can be used to produce a confidence interval for the true, unknown indices.

Keywords: Analysis of variance decomposition; bootstrap; sensitivity indices

1. INTRODUCTION

In the context of numerical experiments, sensitivity analysis (SA) aims to quantify the relative importance of input variables $X = (X_1, \ldots, X_n)$ in determining the value of an assigned output variable $Y = f(X)$. (Note that Y could in fact be vector-valued without affecting any of the following results). More specifically, *global* SA tries to quantify output uncertainty due to the uncertainty in the input variables, both singly and in combination with one another. *Variance based* SA techniques are intended to estimate how much ouput variability is dependent on each of the input variables (again, taken singly and in combination with one another).

This note deals with a recently developed method for global sensitivity analysis of model output: Sobol′ sensitivity indices. The method is based on decomposing the variance of model output into terms of increasing dimensionality, as in the classical Analysis of Variance (ANOVA) of factorial experimental designs. Monte-Carlo (MC) techniques are used to expedite their construction, and a bootstrap method is presented which produces reliable interval estimates for the true, unknown index values. The use of the bootstrap increases the usefulness of the indices by reducing the computational effort required to estimate their variability.

The layout of the paper is as follows. In Section 2 we introduce the Sobol′ indices, and discuss briefly the history of the ANOVA-decomposition, and explain how to construct the indices using MC methods. In Section 3 the bootstrap method for interval estimation is explained, and Section 4 presents a numerical example of all the work. Section 5 contains the conclusions drawn from the experiments, and ideas for future work.

2. THE SOBOL′ SENSITIVITY INDICES

2.1. Motivation

The sensitivity indices described in this note were developed by Sobol′ (1990a), based on his earlier work on the Fourier Haar series (1969). The indices were developed for the purpose of Sensitivity Analysis (SA), that is, to estimate the sensitivity of a function $f(\boldsymbol{X})$ with respect to different variables or subgroups of variables. In SA terminology, $\boldsymbol{Y}=f(\boldsymbol{X})$ is the (possibly vector valued) output variable, while the $\boldsymbol{X}$ are the input variables. The method is outlined briefly:
Let the function $Y=f(X)=f(X_1 \ldots X_n)$ be defined on the n-dimensional unit cube

$$K^n = \{X : 0 \leq X_i ; \leq 1, i = 1, \ldots, n\}$$

It is possible to decompose $f(X)$ into summands of increasing dimensions:

$$f(X_1, \ldots, X_n) = f_0 + \sum_{i=1}^{n} f_i(X_i) + \sum_{1 \leq i < j \leq n} f_{ij}(X_i, X_j) + \cdots + f_{12\cdots n}(X_1, \ldots, X_n) \qquad (1)$$

provided that f_0 is a constant and the integral of every summand over any of its own variables is zero:

$$\int_0^1 f_{i_1 \ldots i_s}(X_{i_1}, \ldots, X_{i_s}) dXi_k = 0, \quad 1 \leq k \leq s. \tag{2}$$

Consequences of (1) and (2) are that all the functions which appear within the summands in (1) are orthogonal, and that $f_0 = \int_{K^n} f(\boldsymbol{X}) d\boldsymbol{X}$. In Sobol′ (1969) the representation (1, 2) is constructed from consideration of Fourier Haar series; in his 1990 article a more general developement is offered. If $f(\boldsymbol{X})$ is integrable in the unit cube, then all of the functions which appear within the summands in (1) are also integrable, as follows:

$$f_i(X_i) = \int_0^1 \ldots \int_0^1 f(\boldsymbol{X}) d\boldsymbol{X}_{-i} - f_0,$$

$$f_{ij}(X_i, X_j) = \int_0^1 \ldots \int_0^1 f(\boldsymbol{X}) d\boldsymbol{X}_{-\{ij\}} - f_i(X_i) - f_j(X_j) - f_0,$$

and so on, where the convention is used that $d\boldsymbol{X}_{-\{ij \ldots m\}}$ indicates integration over all variables with the exception of those within the subscript parenthesis. These integrals will be remarked upon in Section 2.2, where we explore the history of this form of decomposition.

The total variance of $f(\boldsymbol{X})$ can be written as $D = \int_{K^n} f^2(\boldsymbol{X}) d\boldsymbol{X} - f_0^2$,while

$$D_{i_1} \cdots {}_{i_s} = \int_0^1 \ldots \int_0^1 f^2_{i_1 \ldots 1_s}(X_{i_1}, \ldots, X_{i_s}) dX_{i_1} \ldots dX_{i_s} \tag{3}$$

is the contribution to the total variance from term $f_{i_1 \ldots i_s}$in the series development. At this point the sensitivity estimates $S_{i_1 \ldots i_s}$ can be introduced:

$$S_{i_1 \ldots i_s} = \frac{D_{i_1 \ldots i_s}}{D} \tag{4}$$

In Sobol′ (1990a) it is shown that the total variance can be partitioned in the same way as the original function:

$$D = \sum_{i=1}^{n} D_i + \sum_{1 \leq i < j \leq n} D_{ij} + \cdots + D_{12 \ldots n} \tag{5}$$

from which it follows that the sum of all the sensitivity indices-over all possible combinations of indices-must be 1. We write this as $\sum'' S_{i_1 \ldots i_s} = 1$.

This decomposition is useful for SA because the terms $S_{i_1 \ldots i_s}$ give the fraction of the total variance of $f(X)$ which is due to any individual input variable or combination of input variables. In this way, for example, S_1 is the main effect of varaible X_1, S_{12} is the interaction effect, i.e., that part of the output variation due to variables X_1,X_2 which cannot be explained by the sum of the effect of the two variables alone. Finally, the last term $S_{12\ldots n}$ is that fraction of the output variance which cannot be explained by summing terms of lower order.

This decomposition is not unique in the analysis of numerical experiments (see next section): a variance decomposition identical to (5) is suggested by Cukier *et al.* (1978) when using the Fourier Amplitude Sensitivity Test (FAST) method, based on the Fourier transform, for sensitivity analysis. The FAST indices are identical to the Sobol′ ones in all but computation (Saltelli & Bolado, 1996); however their calculation is usually limited to only those indices which refer to "main effect", that is, individual X_i terms (Liepman & Stephanopoulos, 1985).

Decompositions similar to (1) and (4) are discussed in Cotter (1979), Cox (1982), Efron and Stein (1981), and Sacks *et al.* (1989). In this last article the plots of the individual f terms in the development (1) are used for the purpose of SA.

Other investigators (Iman & Hora (1990), Krzykacz (1990), Saltelli *et al.*, (1993), McKay (1995)) have developed sensitivity measures (called importance measures or correlation ratios) which are also based on the fractional contribution of the total variance of individual input variables. In an SA exercise, these measures would produce the same importance ranking as that gained by consideration of the single term S_i values. For a discussion of the various measures, see also Homma & Saltelli (1994).

Practically, in order to apply Sobol′ sensitivity estimates one must evaluate the multidimensional integrals (such as Equation (3)) using MC methods. Each term in the series development (1) is a separate integral, and the number of terms is equal to $2^n - 1$, far too many to be computed even for moderate model dimension n. An MC technique designed to obviate this difficulty – reducing the number of MC calculations to $n+1$- is explained in Section 2.3.

In fractionally replicated designs it is customary to assume that higher order interactions are zero, in order to leave sufficient degrees of freedom for variance estimation, but in SA experiments, where the models are usually nonlinear and the variation in the response much wider, it may happen that the higher order terms are the most important (for example, see Saltelli & Sobol′ (1995)) and so their estimation is crucial. These sensitivity indices allow such effects to be estimated easily, by partitioning $\boldsymbol{X}$, and treating subsets of variables as new variables; for example, $\boldsymbol{X}$ could be partitioned into $(\boldsymbol{U},\boldsymbol{V})$, where $\boldsymbol{U}=(X_1 \ldots X_k)$, and $\boldsymbol{V}(X_{n-k+1,\ldots,}X_n)$. The variance of $f(\boldsymbol{X})$ can then be decomposed into $D=D_U+D_V+D_{UV}$. Using this results, we define a "total effect" term for each variable, by letting $\boldsymbol{U}=X_i$, $\boldsymbol{V}=(X_1,\ldots,X_{i-1},X_{i+1}\ldots X_n)$ and declaring the total effect of variable i to be given by

$$S_{T_i}=S_i+S_{i,v}=1-S_v.$$

In this way the *total contribution of each variable to the output variation* is estimated.

The general conclusion to this subsection is that the Sobol′ formulation of the sensitivity indices is very general and includes as a particular case most of what has been done previously in SA, using decompositions like (3) or (5), as well as those sensitivity measures based on fractional contributions to the output variance. Sobol′ indices of the first order are identical to FAST coefficients, and differ from the other tests mentioned above only by a scale factor. Further, Sobol′ indices allow the interaction and higher order interaction terms to be computed straightforwardly; this capacity makes Sobol′ sensitivity analysis similar to the Analysis of Variance in factorial design, as shall be discussed in the next section.

2.2. The ANOVA Decomposition

The decomposition (1) or (5) has a long history which, in this section, is examined, to help endow the Sobol′ indices with their true statistical definitions. The decomposition has an interesting pedigree which links naturally to U-statistics and resampling methodology. The lemma which describes the decomposition in its most general form is given in Efron & Stein (1981), and concerns functions of independent random variables $\{X_i\}$ which are not necessarily identically distributed, although when they are, the result can be stated more concisely. Suppose, as above, $f(X_1, \ldots, X_n)$ is some statistic defined on the product measure generated by $\boldsymbol{X} = (X_1, \ldots, X_n)$. Then $f(\boldsymbol{X})$ may be decomposed into a grand mean $f_0 = E[f(\boldsymbol{X})]$, i'th main effect $f_i(X_i) = E[f(\boldsymbol{X}|X_i = x_i)] - f_0$; ij'th *interaction* $f_{ij}(X_i, X_j) = E[f(\boldsymbol{X})| X_i = x_i, X_j = x_j] - E[f(\boldsymbol{X})|X_i = x_i] - E[f(\boldsymbol{X})|X_j = x_j] + f_0$, and so on. Given these definitions, the decomposition in Section 2.1 follows immediately, as the case $n = 2$ easily demonstrates:

$$\begin{aligned} f(X_1, X_2) &= f_0 + f_1(X_1) + f_2(X_2) + f_{12}(X_1, X_2) \\ &= f_0 + E[f(\boldsymbol{X})|X_1 = x_1] - f_0 + E[f(\boldsymbol{X})|X_2 = x_2] - f_0 \\ &\quad + E\,[f(\boldsymbol{X})|X_1 = x_1, X_2 = x_2] - f_0 \\ &\quad - E[f(\boldsymbol{X})|X_1 = x_1] - E[f(\boldsymbol{X})|X_2 = x_2] + f_0 \\ &= f(X_1, X_2). \end{aligned}$$

Using the law of iterated expectations. it is straightforward to see all the random variables on the right hand side have zero mean and are mutually uncorrelated. For example $E[f_i(X_i)] = E[E\{f(\boldsymbol{X}|X_i = x_i) - f_0\}] = E[f(\boldsymbol{X})] - f_0 = 0.$ □

This decomposition of a statistic is the same as is typically deployed for data collected from a factorial experiment (Fisher, 1958). One of our aims is to discuss the similarities and the differences between SA and ANOVA.

One of the earliest to write about this lemma was Hoeffding (1948), who was concerned with estimators $f(\boldsymbol{X})$ which are U-statistics. Such a U-statistic is defined as

$$U = \frac{\sum f(X_{\alpha_1}, \ldots, X_{\alpha_m})}{n(n-1)(n-2)\ldots(n-m+1)},$$

where, as in Section 2.1, the sum in the numerator is carried out over all permutations $(\alpha_1, \ldots, \alpha_m)$ of m different integers. If the $\{X_i\}$ are independent and identically distributed (i.i.d.) according to some distribution function G, then U is an unbiased estimator of

$\theta(G) = \int \cdots \int f(x_1, \ldots, x_m)dG(x_1) \ldots dG(x_m)$. In other words U is the average of all the possible values of $f(X_{\alpha_1}, \ldots, X_{\alpha_m})$ drawn from the realisation of $(X_1, \ldots, X_n), m < n$. This will strike chords with readers who are practioners of resampling methodologies (start with Efron 1979), which have become steadily more popular as cheap and powerful computing facilities become more widely available. In bootstrapping, resamples are drawn from the full n set of data $(x_1, \ldots, x_n)$, while the U-statistics are even more redolent of the jackknife (Chapter 5 of Efron & Tibshirani (1993)), which estimates functionals of G by forming a weighted average of the set of estimates obtained by deleting one (or more) data points at a time.

Both the jackknife and the bootstrap use as a rationale the substitution of the sample distribution function $\hat{G}$ (which places probability mass $1/n$ on each x_i) for the unknown population distribution G. Substituting in the function $\theta(G)$ above gives

$$\theta(\hat{G}) = \frac{1}{n^m} \sum_{\alpha_1=1}^{n} \cdots \sum_{\alpha_m=1}^{n} f(X_{\alpha_1}, \ldots, X_{\alpha_m}).$$

Hoeffding shows that $\theta(\hat{G})$ is a linear function of U-Statistics with $E[\theta(\hat{G})] = \theta(G) + O(n^{-1})$ and therefore $\theta(\hat{G}) \rightarrow \theta(G)$ as $n \rightarrow \infty$. This fact-that as sample size increases, so an estimator calculated on a random sample will tend towards the population parameter-justifies the bootstrap also. Interestingly enough, if f is a linear statistic, i.e., all terms on the right hand side of the decomposition are zero apart from those which only involve individual X_i, then the jackknife and bootstrap estimates of θ agree, whilst if there are higher-order effects then the bootstrap estimates are more accurate.

Much of the rest of Hoeffding's rich paper is a mathematical *tour-de-force*, in which it is shown that many commonly used statistics, such are the rank correlation coefficient, are examples of U-statistics. The most important result is that when the $\{X_i\}$ are i.i.d. and f is a plug-in estimator (that is, it does not depend on n), then asymptotically $\sqrt{n}(U - \theta)$ has a normal distribution.

The relationship between U-statistics and the jackknife is explored in Efron & Stein (1981), who consider the case where

$$E_G(X_i) = \zeta \quad \text{and} \quad \text{var}_G(X_i) = \sigma^2 \forall i$$

and show that the plug-in estimator of sample variance

$$S(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2, \quad \text{where} \quad \bar{X} = \sum_{i=1}^{n} X_i$$

has grand mean: $f_0 = (\frac{n-1}{n})\sigma^2$,
main effects: $f_i(X_i) = \frac{n-1}{n^2}[(x_i - \zeta)^2 - \sigma^2], \forall i$,
and pairwise interaction effects: $f_{ij}(X_i, X_j) = \frac{1}{n^2}(x_i - \zeta)\ (x_j - \zeta), \forall(i,j)$
and all higher order terms are zero.

A more interesting application to SA is the realisation that the decomposition (1), which powers the Sobol′ indices, is exactly the same as that used to construct an ANOVA of the results of a factorial experimentation. In this sense, ANOVA is SA, and SA is ANOVA. For consider a response variable X, measured under different conditions of two factors A (with I levels) and B (with J levels). Each particular combination of factors is replicated K times, and so X_{ijk} is the k'th replicate of the experiment at the (I, J)'th level of factors A and B. (Attention is restricted to the case of the fully replicated, two factor design only for ease of discussion). In statistical terminology, A and B are factors, X is a response variable; in SA, A and B are input variables, and X is the output variable. But the Sobol′ indices, to measure the importance of A, B, or their interaction (written $A\ B$) uses exactly the same decomposition as the ANOVA F-tests of significance. This is seen if we write down the total sum of squares about the mean, used in ANOVA to perform the F-test. We have:

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - X_{...})^2 = JK\sum_i(X_{i..} - X_{...})^2 + IK\sum_j(X_{.j.} - X_{...})^2 \\ + \sum_i\sum_j(X_{ij.} - X_{i..} - X_{.j.} + X_{...})^2 \\ + \sum_i\sum_j\sum_k(X_{ijk} - X_{ij.})^2\end{aligned}$$

(A dot in the subscript indicates that the average has been taken over that index.) In comparison with (5), the first two terms on the right hand side correspond to the single D terms ($N=2$ here), with the third right hand side term corresponding to D_{12}. The final term on the right hand side is the residual sum of squares-it has expected value of zero and is used to measure the deviance of the data from the theoretical normal linear model. Coupling this decomposition with a Gaussian assumption about the X_{ijk} allows the significance testing of the various effects (factors, or input variables). For example, to test the strength of factor A on X, the ratio of the first term on the left hand side to the residual sum of squares is constructed, and compared to the appropriate F distribution.

These ANOVAs were first employed (Fisher, 1958) in biological experiments however, where random variation-here modelled through $\{\varepsilon_{ijk}\}$, and estimated with the residual sum of squares-has to be taken into account. Is this the case in SA? Sacks *et al.* (1989) have argued that, in a computer experiment, all the variation in the response comes through the variation in the input variables and so models of the above form are inappropriate. This would certainly be the case if only two input variables were under study. But as Terry Andres has pointed out (private communication) in a typical computer experiment the statistician is faced with usually at least dozens of input variables. In that case, one could fit a model using $m < n$ of them, and test its adequacy by assuming that the residual noise in the variation of the input constitutes random error. This is in fact exactly what happens in "classical" ANOVA: the only difference being that for a computer experiment we are aware of the causes of the extra noise, while in the biological experiment we are not. Although conceptually different to fractionally replicated designs-where higher order interactions are set zero in order to leave sufficient degrees of freedom with which to estimate σ^2-the approach is identical. The fact that investigators in different disciplines have used the same kind of decomposition, apparently without cross-fertilisation, seems to be one of the many instances of convergent thinking in the handling of scientific problems.

2.3. Computation of Sobol′ Sensitivity Indices

As mentioned in the previous section Sobol′ indices can be computed using plain Monte Carlo integrals. So, for example, if N sets of X are generated, $x_1, x_2, \ldots, x_N$ say, each one a sampled point in K^n, then the straightforward MC estimates of grand mean and total variance are given by

$$\hat{f}_0 = \frac{1}{N}\sum_{m=1}^{N} f(\boldsymbol{x}_m); \quad \hat{D} = \frac{1}{N}\sum_{m=1}^{N} f^2(\boldsymbol{x}_m) + \hat{f}_0^2$$

However, it is not necessary to generate a further set of MC samples for *each* combination of variables whose sensitivity index must be calculated. If we write

$$\boldsymbol{x}_{-i,m} = (x_{1m}, \dots, x_{i-1,m}, x_{i+1,m}, \dots, x_{nm})$$

then at the *m*'th stage of the MC process we generate two such samples, $\boldsymbol{x}_{-i,m}^{(1)}$ and $\boldsymbol{x}_{-i,m}^{(2)}$ say, and then the partial variances required to estimate the main effects ($S_i = D_i/D$) can be estimated using

$$\hat{D}_i = \frac{1}{N}\sum_{m=1}^{N} f(\boldsymbol{x}_{-i,m}^{(1)}, x_{im}) f(\boldsymbol{x}_{-i,m}^{(2)}, x_{im}) - \hat{f}_0^2$$

here, the superscripts (1) and (2) refer to different MC samples. The computational procedure may be clarified as follows:

(i) Under the assumption that all main effects are to be estimated, two matrices of dimension N by n are generated (matrices (1) and (2) in the previous equation). The matrix superscripted (1) is used for "sampling" and that superscripted (2) is for "re-sampling" (although not in the bootstrap sense!).

(ii) In order to compute variable X_i's contribution to the total variance, multiply the values of f obtained by sampling independently all the variables by the corresponding f values obtained by "resampling" all the variables except X_i. If variable i is important, then high values of the first term in the ($f \times f$) product will be multiplied by similarly high values in the second term. Otherwise the pairing of terms will tend to be casual, high values being possibly multiplied by low ones, and the $\hat{D}_i$ value tend to be lower. This technique was proposed in Saltelli *et al.* (1993) but had already been described in a Russian article (Sobol′ (1990a); more convergent thinking). The approach aims to minimise MC variability, since sensitivity indices for variables i

and j will differ only in the i'th and j'th columns of their input matrices, S_i is in fact computed by summing terms:

$$f(x^{(1)}_{-ijm}\, x^{(1)}_{im}\, x^{(1)}_{jm}) \times f(x^{(2)}_{-ijm}\, x^{(1)}_{im}\, x^{(2)}_{jm})$$

while S_j from summing terms:

$$f(x^{(1)}_{-ijm}\, x^{(1)}_{im}\, x^{(1)}_{jm}) \times f(x^{(2)}_{-ijm}\, x^{(2)}_{im}\, x^{(1)}_{jm})$$

(In computing S_i we resample all but x_i, while for S_j we resample everything except x_j.)

(iii) The computational cost of the procedure, the number of model evaluations required to compute all the main effects, is size $(n+1)N$. (One sample of size N for computing the average $\hat{f}_0$ plus n of the same size for each of the main effects). The cost of the bootstrap procedure in Section 3 is negligible by comparison, as the bootstrap uses the model evaluations already generated. In normal test applications, it is more expensive to generate a single f value than to resample a set of f values a large (10,000) number times.

(iv) The higher order interactions and the total effects terms are computed in the same manner. For instance the *total* effect for X_i (see Section 2.1) requires the calculation of

$$\hat{D}_{T_i} + \hat{f}_0^2 = \frac{1}{N}\sum_{m=1}^{N} f(\mathbf{x}^{(1)}_{-im}, x^{(1)}_{im}) f(\mathbf{x}^{(1)}_{-im}, x^{(2)}_{im}), \tag{6}$$

where this time only one realisation $\mathbf{x}_{im}$ is "resampled" to obtain the second term in the product. Therefore, in order to compute a full set of total effect terms $\{\hat{S}_{T_i}\}$, it is only necessary to compute one Monte Carlo integral for the mean, plus n MC integrals for the $\{\hat{D}_{T_i}\}$.

2.3.1. Sample Generation

In his work on sensitivity measures for nonlinear models, Sobol′ recommends the use of a quasi random numbers sequence for the computation of the MC integrals: the LP_τ sequence. A description of this algorithm can be found in Sobol′ (1990b); quasi random numbers are characterised by faster convergence under certain limitations on the value of n; see also Davis & Rabinowitz (1984), and Brately & Fox (1988).

3. BOOTSTRAP CONFIDENCE INTERVALS FOR SENSITIVITY INDICES

No estimate of sensitivity can be of any use without an estimate of its sampling variability. In Sobol′ (1990b) the "probable error" for the D_i was used as a judge of accuracy, i.e., δD_i is calculated so that $Pr\{|D_i - \hat{D}_i| \leq \delta D_i\} = 0.5$. For the S_i and S_{T_i} indices, a more suitable estimate of accuracy can be derived using bootstrap confidence intervals (BCIs). For detailed discussion of BCIs, see Chapters 12,13,14 and 22 of Efron & Tibshirani (1993); basically, the MC sampled $\{\boldsymbol{x}_{im}\}$ values are resampled (i.e., sampled with replacement) B times, at each stage and for each variable S_i is recalculated, leading to a bootstrap estimate of the sampling distribution of the sensitivity indices, $\{S_i^{*b}\}_{b=1}^{B}$, for $i = 1, \ldots, n$. Then BCIs can be constructed in a number of ways: as a small check on robustness in our numerical work, two methods were used. First, the percentile method, which selects as endpoints for a 95% interval the 2.5% and 97.5% percentiles of the bootstrap distribution. Secondly, the moment method, which relies on large sample theory and gives a symmetric 95% interval for S_i as

$$\hat{S}_i \pm 1.96 \times e.s.e.\ (\hat{S}_i), \text{where } e.s.e.(\hat{S}_i) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(S_i^{*b} - S_i^{*\bullet})^2}$$

$$\text{and } S_i^{*\bullet} = \frac{1}{B}\sum_{b=1}^{B} S_i^{*b}.$$

In our experiments, we chose $B = 10000$. The moment interval has the advantage that the bootstrap provides reliable estimates of standard error for very many fewer sizes of B than is the case for percentiles of the unknown distribution function. However, moment intervals could have poor coverage properties if the distribution turned out to be skewed either to the right or the left.

Bootstrapping works because sampling with replacement from a set of independent, identically distributed data is equivalent to sampling from the empirical distribution function of the data. In Appendix A it is shown that the procedure, usually applied with pseudo-random sampling, is valid with the use of quasi-random sampling as well.

3.1. A Note on the Number of Resamples

One of our reviewers comments that the bootstrap technique may be offputting due to the large number of resamples required. Is $B = 10{,}000$ a prohibitive number, for a practical, as opposed to artificial, example? In the case of independent, identically distributed data, the time needed to calculate each S_i^* really depends on the length of time required to select N random variates uniformly distributed on (0,1). This is not likely to take a long time (our total computing time for both sets of confidence intervals was about 5 minutes). It is the case that reliable estimation of the extreme percentiles of a sampling distribution requires much larger amounts of resampling than is the case for simpler standard error estimation (Hall, 1986). We re-iterate for emphasis: bootstrap resampling is carried out *without the need for further model evaluations*. It is the number of model evaluations which govern the cost of a numerical example.

In practice a value of $B = 1000$ or 2000 is likely to be chosen. A very simple check on the adequacy of the percentile intervals can be given by the *jackknife-after-bootstrap* method (Efron 1992). Without any further resampling, this technique estimates the standard errors in bootstrap functionals, and so a quick check can be made on the stability of bootstrap intervals generated from (say) $B = 1000$. If the percentiles seem robust against bootstrap sampling variability, then the experimenter can be happy with a much lower level of resampling than used in our experimental section.

4. NUMERICAL EXPERIMENTS

The effectiveness of the Sobol′ indices and the bootstrap intervals are investigated using the following test function:

$$f(X) = \prod_{i=1}^{n} g_i(X_i),$$

where

$$g_i(X_i) = \frac{|4X_i - 2| + a_i}{1 + a_i}, \quad a_i \geq 0$$

As before, f is defined on K^n. Figure 4.1 gives plots of the function $g_i(\bullet)$, for different values of the parameter a_i. Function f will integrate to 1 for all values of the parameter greater than or equal to zero.

This test function, with $a_i = 0 \forall i$, was used in Davis & Rabinowitz (1984) to test multidimensional integration. The function $g_i(\bullet)$ was also used in Saltelli & Sobol (1995a, 1995b).

The parameters control the function as follows:

$$1 - \frac{1}{1 + a_i} \leq g_i \leq 1 + \frac{1}{1 - a_i}.$$

The importance of input variable X_i is therefore governed by the size

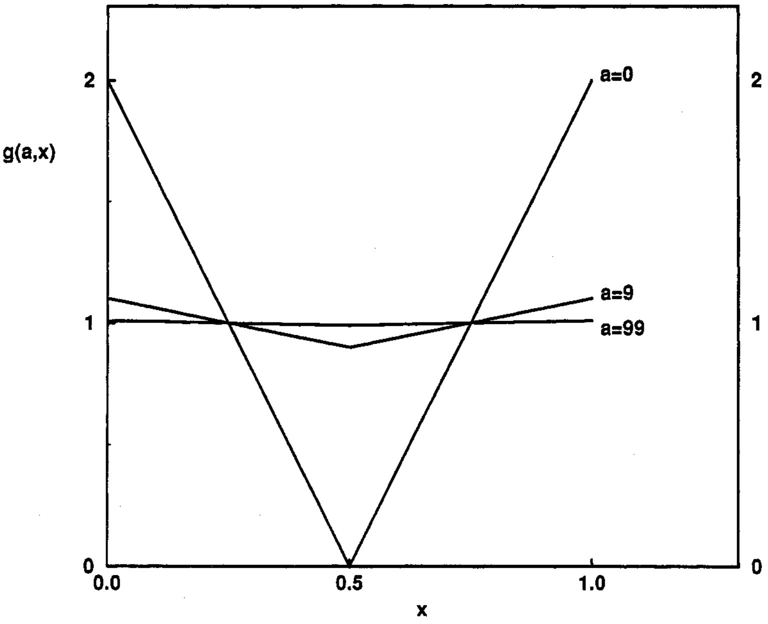

FIGURE 4.1 Some examples of the test function.

of a_i. For example:

$a_i = 0 \Rightarrow 0 \leq g_i \leq 2 \Rightarrow X_i$ is an important variable
$a_i = 9 \Rightarrow 0.9 \leq g_i \leq 1.1 \Rightarrow X_i$ is less important
$a_i = 99 \Rightarrow 0.99 \leq g_i \leq 1.01 \Rightarrow X_i$ is insignificant.

There are several advantages to the use of this test function:

(1) the sensitivity indices can be calculated analytically;
(2) it is strongly nonlinear and nonmonotonic, and hence provides a good test of the methods;
(3) by definition, all the interaction terms are non-zero;
(4) the parameters of the function allow the experimenter to "fine-tune" the level of difficulty.

In addition, the results from this test function are easily understood and comparable with other work (for example in David & Rabinowitz (*ibid.*)). None of these advantages are present in the case of numerical simulation of a numerical experiment. (For the application of the sensitivity indices to larger models, see Saltelli *et al.*, (1993), Saltelli & Hjorth (1995) and SAMO (1996).)

For the experiment, $n = 20$ variables were chosen and the parameters set to $a_i = ((i-1)/2)$ $i = 1, \ldots, n$, and so X_i monotonically decreases in importance as i increases. The input variables were generated using Sobol′ LP_τ sequences, with $N = 512$.

The global indices, $\hat{S}_{T_i}$, were then calculated as described in Section 3. Table IV.I contains a list of the variable rankings, estimated sensitivities, and their interval estimates. Figures 4.2a and 4.2b contain the same information graphically.

5. RESULTS

Starting with the indices themselves, there is good monotonic decrease in importance for the first seven variables. The estimates for Variables 11,12 and 18 are a little odd, but the indices have indeed picked out the most important variables clearly. Looking at variables 19 and 20, we see the "impossible" result of a negative index. Theoretically impossible, since the indices are theoretically a ratio of variances, it must be concluded that this is a function of the MC "short-cut"-examination of Equation (6) shows how it can occur. However it does

TABLE IV.I Estimates of Sobol′ indices and Bootstrap Confidence Intervals

Variable	$\hat{S}_i$	*Moment BCI*	*Percentile BCI*
1	0.496	(0.390,0.604)	(0.389,0.603)
2	0.244	(0.111,0.378)	(0.104,0.373)
3	0.134	(0.045,0.225)	(0.042,0.225)
4	0.062	(−0.009,0.131)	(−0.010,0.132)
5	0.065	(−0.005,0.135)	(−0.009,0.132)
6	0.042	(−0.018,0.101)	(−0.022,0.099)
7	0.038	(0.002,0.072)	(0.000,0.072)
8	0.006	(−0.050,0.063)	(−0.052,0.063)
9	0.025	(−0.016,0.065)	(−0.016,0.065)
10	0.002	(−0.048,0.052)	(−0.049,0.053)
11	0.013	(−0.031,0.057)	(−0.034,0.056)
12	0.015	(−0.018,0.049)	(−0.019,0.048)
13	0.005	(−0.026,0.036)	(−0.026,0.037)
14	0.005	(−0.027,0.036)	(−0.027,0.036)
15	0.001	(−0.029,0.031)	(−0.030,0.030)
16	0.007	(−0.020,0.034)	(−0.020,0.033)
17	0.008	(−0.012,0.028)	(−0.012,0.028)
18	0.014	(−0.009,0.035)	(−0.009,0.035)
19	−0.010	(−0.032,0.012)	(−0.032,0.013)
20	−0.020	(−0.039,−0.001)	(−0.039,0.000)

not affect the message of the results with respect to the relative importance of the input variable.

Examining the BCIs, there is very close agreement between the two methods, and so it seems very likely that the sampling distribution of these indices is symmetrical, a hypothesis which gains extra weight when it is noted that the index estimates are by and large near the centre of the BCIs. A "quick-and-dirty" hypothesis test of index differences using the intervals suggests that the importance measure for Variable 2 is significantly lower than that for variable 1 (since 0.244 lies below the lower endpoint of the Variable 1 intervals). Indeed Variable 1 in such a comparison appears significantly more important than any other variable. More is said on this matter of index differences in Section 6.

Another approach common to SA is to rank transform data to make analysis more "robust" (see Iman & Conover (1979), Saltelli & Marivoet (1990) and Saltelli *et al.*, (1993)). The Sobol′ indices are certainly amenable to rank transformation and in fact using the test function of Section 4 on rank transformed input variables served to emphasise the difference between the first and subsequent variables-with the pleasing results that the negative indice estimates for the very

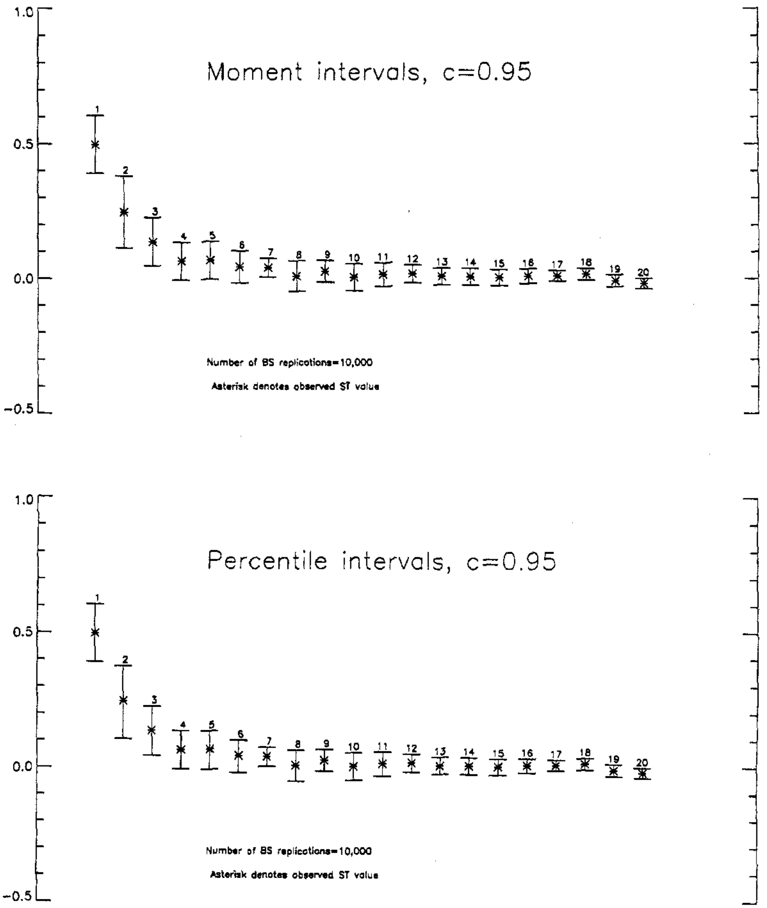

FIGURE 4.2 Moment (top) and Percentile (bottom) bootstrapped intervals for the total Sensitivity Indices.

unimportant variables disappeared. The inference about relative variable importances was unaffected by the transformation and so the results are not displayed.

6. CONCLUSIONS

We have offered a short description of a new, powerful method for sensitivity analysis, which incorporates advantages from several different pre-existing approaches (from FAST to correlation ratios). It builds on the earlier methods by allowing a full ANOVA to be carried out on the outcome of a numerical experiment. Results from the literature and the example in Section 4 are certainly encouraging. The Sobol′ sensitivity indices were able to clearly "sort out" the most important input variables into their correct order of importance, and the bootstrap intervals were an effective way of assessing overlap between them.

The bootstrap has been shown useful in the situation of gauging uncertainty in an important area-where uncertainty itself is under examination! However, it should be recognised that comparing confidence intervals across classes is not a formal test of the hypothesis that one sensitivity index is significantly different from another, but only a "quick" method of assessing how likely this is to be the case. It would be conceptually straightforward to use the bootstrap to test such a hypothesis: and clearly it is of interest so to do. The bootstrap could also be used to examine the bias in an SA estimate, and if necessary, correct that bias. This exciting idea calls for further development.

The global effect indices (those which include all the effects of a parameter, either alone (the main effect) or in conjunction with others (the interaction) are especially powerful for SA, particuarly when used with the bootstrap estimates of confidence bounds. By taking into account higher order interactions, the indices seem to offer a useful advantage over previously advanced variance based SA measures.

The use of the indices, exemplified here through their application to a well known analytic test case, is especially useful for real applications. In Saltelli *et al.* (1993) it was shown, via reference to three international benchmarks, that variance based measures are preferrable for automated sensitivity analysis. In Saltelli & Hjorth (1995) the global effects indices were applied successfully to a chemical kinetics system. In the same article, it was also shown that the interaction terms are indeed important in many real settings, whereby a first order sensitivity measure may be insufficient. However, Saltelli and Hjorth (1995) estimated the error on the sensitivity measure using the probable error (see Section 3 above). The introduction of the bootstrap approach in this article upgrades the performance of the measures by offering a more accurate and realistic evaluation of the error. This improvement partially mitigates the main drawback in the use of the indices, i.e., the large sample size $((n+1)N)$ needed for their computation; this is also a consequence of the MC approach applied to a continuous range of x values. As discussed in Section 2, a (saturated or fractional) factorial design model could be used in the same problem setting, and future work could certainly involve a comparison of Sobol′ and ANOVA approaches.

References

[1] Bratley, P. and Fox, B. L. (1988). Algorithm 659. Implementing Sobol′s quais-random sequence generator. *ACM Transactions on Mathematical Software*, **14**, 88 – 100.

[2] Cotter, S. C. (1979). A screening design for factorial experiments with interactions. *Biometrika*, **66**, 317 – 320.

[3] Cox, D. C. (1982). An analytical method for uncertainty analysis of nonlinear output functions, with applications to Fault-tree analysis. *IEEE Transactions on Reliability*, **R-31**, 465 – 468.

[4] Cukier, R. I., Fortuin, C. M., Schuler, K. E., Petschek, A. G. and Schaibly, J. K. (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. *J. Chem. Phys.*, **59**, 3873 – 3878.

[5] Davis, P. J. and Rabinowitz, P. (1984). *Methods of numerical integration* (2nd edition), Academic Press, New York.

[6] Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions (with discussion). *Journal of the Royal Statistical Society, series B*, **54**, 83 – 127.

[7] Efron, B. and Stein, C. (1981). The Jackknife estimate of variance, *Annals of Statistics*, **9**, 586 – 596.

[8] Efron and Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

[9] Fisher, R. A. (1958). *Statistical Methods for Research Workers* (13th edition), Oliver and Boyd, Edinburgh.

[10] Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, **14**, 1453 – 1462.

[11] Hoefding, W. F. (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, **19**, 293 – 325.

[12] Iman, R. L. and Conover, W. J. (1979). The use of rank trasform in regression. *Technometrics*, **21**, 499 – 509.

[13] Iman, R. L. and Hora, S. C. (1990). A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Analysis*, **10**, 401 – 406.

[14] Krzykacz, B. (1990). SAMOS: A computer program for the derivation of empirical sensitivity measures of results from large computer models. *Technical Report* **GRS-A-1700**, Gesellschaft fuer Reaktor Sicherheit (GRS), MbH, Garching, Germany.

[15] Liepman, D. and Stephanopoulos, G. (1985). Development and global sensitivity analysis of a closed ecosystem model. *Ecological Modelling*, **30**, 13 – 47.

[16] McKay, M. D. (1995). Evaluating Prediction Uncertainty. *Los Alamos National Laborator Technical Report* **NUREG/GE-6311**.

[17] Sacks, J., Welch, S. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, **4**, 409 – 435.

[18] Saltelli, A., Andres, T. H. and Homma, T. (1993). Sensitivity Analysis of model output: an investigation of new techniques. *Computational Statistics and Data Analysis*, **15**, 211–238.

[19] Saltelli, A. and Bolado, R. (1996). Is there another way to compute Fourier Amplitude Sensitivity Test (FAST)? Submitted to *Computational Statistics and Data Analysis*.

[20] Saltelli, A. and Hjorth, J. (1995). Uncertainty and sensitivity analyses of OH-initiated dimethylsulphide (DMS) oxidation and kinetics. *J. Atmospheric Chemistry*, **21**, 187–221.

[21] Saltelli, A. and Homma, T. (1992). Sensitivity analysis for model output. Performance of balckbox techniques on three international benchmark exercises. *Computational Statistics and Data Analysis*, **13**.

[22] Saltelli, A. and Marivoet, J. (1990). Nonparametric statistics in sensitivity analysis for model output; A comparison of selected techniques. *Rel. Eng. and System Safety*, **28**, 229–253.

[23] Saltelli, A. and Sobol′, I. M. (1995a). Sensitivity analysis for nonlinear mathematical models: numerical experience. *Matematicheskoye Modelirovaniye*, **7**, 16–29. (In Russian).

[24] Saltelli, A. and Sobol′, I. M. (1995b). About the use of rank transformation in sensitivity analysis of model output. *Reliability Engineering and Systems Safety*, **50**, 225–239.

[25] SAMO (1996). Proceedings of the Symposium on Theory and Applications of Sensitivity Analysis of Model Output in Computer Simulation (SAMO), Belgirate, Italy, September 1995. EUR Report 16331, ISBN 92-827-5530-4. Editors: A. Saltelli and H. Von Maravic', Luxembourg 1996.

[26] Sobol′, I. M. (1969). *Multidimensional quadrature formulas and Haar functions*, Nauka, Moscow. (In Russian).

[27] Sobol′, I. M. (1990a). Sensitivity estimates for nonlinear mathematical models, *Matematicheskoe Modelirovanie*, **2**, 112–118 (in Russian), translated in *Mathematical Modelling and Computational Experiments*, **1**, 407–414 (1993).

[28] Sobol′, I. M. (1990b). Quasi-Monte Carlo methods. *Progress in Nuclear Energy*, **24** 55–61.

[29] Sobol′, I. M. (1994). *A Primer for the Monte Carlo Method.* CRC Press, Boca Raton.

[30] Sobol′, I. M. and Shukhman, B. V. (1995). Integration with quasirandom sequences: numerical experiments. *J. of Modern Phys. C.*, 263–275.

APPENDIX A QUASI-RANDOM SAMPLING FOR BOOTSTRAP

Suppose θ is a functional of the distribution F which generated the i.i.d. sample $\boldsymbol{x}$ ($F \rightarrow \boldsymbol{x}$), and that $\hat{\theta} = s(\boldsymbol{x})$ is the estimator of this functional. The bootstrap replicate $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)$ is generated by sampling *at random and with replacement* from $\boldsymbol{x}$.

The simplest algorithm for sampling $\boldsymbol{x}^*$ on a computer can be specified (Sobol′, 1994) as follows: select a standard random number γ, compute an integer $k = [n\gamma] + 1$ and put $x^* = x_k$. This rule defines a transformation $x = h(\gamma)$, and the function $h(z)$ is piecewise constant.

So we can write $\boldsymbol{x}^* = (h(\gamma_1), \ldots, h(\gamma_n))$, and see that the resample is defined by one random point Γ with cartesian co-ordinates $(\gamma_1, \ldots, \gamma_n)$. Clearly, Γ is distributed in the n-dimensional unit hypercube.

Bootstrap Sample

The bootstrap distribution for $\hat{\theta}$ is generated as follows. Select B independent random points $\Gamma_b, b = 1, \ldots, B$, from which are generated B resamples, each of which yields a re-estimate of the functional of interest:

$$\theta^{*b} = s(\boldsymbol{x}^{*b}), \quad b = 1, \ldots, B$$

which is the bootstrap estimate of the sampling distribution. It can be used to estimate various moments of the estimator. Estimates of the expectation and standard error are, respectively,

$$\theta_B^{*\bullet} = \frac{1}{B}\sum_{b=1}^{B} \theta^{*b}, \quad se_B^2 = \frac{B}{B-1}\left[\frac{1}{B}\sum_{b=1}^{B}(\theta^{*b})^2 - (\theta_B^{*\bullet})^2\right].$$

At large B, $\theta_B^{*\bullet}$ is asymptotically normal and the definition of approximate probable errors or confidence intervals is possible. Both quantities stochastically converge:

$$\theta_B^{*\bullet} \to E[\hat{\theta}], se_B^2 \to E[\hat{\theta}^2] - (E[\hat{\theta}]^2), \quad \text{as} \quad B \to \infty,$$

where

$$E[\hat{\theta}^m] = \int_0^1 \cdots \int_0^1 [s(h(z_1), \ldots, h(z_n))]^m dz_1 \ldots dz_n.$$

The integral can be easily expressed in the form of a sum because the integrand is piecewise constant:

$$E[\hat{\theta}^m] = \frac{1}{n^n}\sum_{i_1=1}^{n} \cdots \sum_{i_n=1}^{n} [s(x_{i_1}, \ldots, x_{i_n})]^m.$$

However, the last expression is as a rule impractical: the number of summands n^n is usually very large.

Quasi-random Bootstrap

Let $Q_1, \dots, Q_b, \dots$ be an n-dimensional quasi-random sequence (that is, a sequence of nonrandom points uniformly distributed in the number-theoretical sense inside the n-dimensional unit hypercube). The bootstrap replications can be computed from points Q_b rather than Γ_b, using the cartesian co-ordinates $(q_1, \dots, q_n)$ of a point Q to generate the resample $\boldsymbol{x}^*$, instead of the random co-ordinates $(\gamma_1, \dots, \gamma_n)$ of Γ. We label the b'th set of quasi-random co-ordinates as $(q_1^b, \dots, q_n^b)$.

Clearly, the function $s(h(z_1), \dots, h(z_n))$ is Riemann-integrable Therefore

$$\frac{1}{B}\sum_{b=1}^{B} s(h(q_1^b), \dots, h(q_n^b)) \to \int_0^1 \cdots \int_0^1 s(h(z_1), \dots, h(z_n)), \text{ as } B \to \infty$$

and the same is true for all positive powers of s. Hence

$$\theta_B^{*\bullet} \to E[\hat{\theta}],\ se_B^2 \to E[\hat{\theta}^2] - (E[\hat{\theta}])^2, \quad \text{as } B \to \infty.$$

The function $\hat{\theta}$ will generally be symmetric (see above) and so (Sobol′ and Shukhman, 1995) switching to quasi-random numbers may lead to a considerable increase in rate of convergence, for moderate n, say $n \leq 15$. At higher n remians only a gain in reliability of computations.