Input Screening: Finding the Important Inputs on a Budget

Max D. Morris

Departments of Statistics, and Industrial and Manufacturing Systems Engineering
Iowa State University, Ames, Iowa, 50011, USA
E-mail: mmorris@iastate.edu

Abstract: One general goal of sensitivity or uncertainty analysis is the determination of which inputs most influence model outputs of interest. Simple methodologies based on randomly sampled input values are attractive because they require few assumptions about the nature of the model. However, when the number of inputs is large and the computational effort required per model evaluation is significant, techniques based on more complex assumptions, analysis techniques, and/or sampling plans are needed. This talk will review some approaches that have been proposed for input screening, with an emphasis on the balance between assumptions and economy, including a brief description of recent work in economical sampling plans.

Keywords: Computer experiment, sensitivity analysis, uncertainty analysis

1. INTRODUCTION

Especially in the early stages of work with a computer model, it is important to determine which inputs are *important* and which are not. The precise definition of "important" is not always the same (and in some cases is never carefully addressed) but is generally related to how much or what kind of influence each input has on outputs of interest. For very simple computer models, such questions may be addressed directly through analysis of the underlying equations. But more complex models require an empirical approach, or *computer experiment* designed to allow determination of the importance of inputs through analysis of numerical output values. The approaches we shall discuss are described as entirely empirical (i.e. "black box"), even though it is understood that in many applications these can be tailored to take advantage of specific knowledge about a model.

In order to be specific, let $y = m(\mathbf{x}), x \in \Delta$, represent what we mean by a "computer model", a deterministic function mapping a vector $\mathbf{x}$ of k input arguments from a defined domain Δ to a scalar-valued output y. In most real problems y would also be vector-valued, but we shall not address complications that this may created here. A particular input x_i may be deemed important if (1.) $\partial y/\partial x_i$ is large in at least some regions of Δ, (2.) y is relatively complex (in some sense) as a function of x_i, or (3.) y varies substantially as the value of x_i changes. These three concepts of "importance" are relatively vague, certainly related, and certainly not exclusive, but one or more of them have been found to be useful in a large variety of problems.

Two characteristics of this problem that make identifying important inputs practically difficult are (1.) the dimension of $\mathbf{x}$ (typically not small), and (2.) the effort required to evaluate m (typically not trivial). The difficulty is easy to understand; if k is large, the number of "points" needed to "fill" it sufficiently to allow characterization of y as a function of $\mathbf{x}$, without extensive knowlege or assumptions about the nature of m, will also be large. But computer experiments requiring a large number of model executions will be prohibitive if each execution is expensive.

Methodologies for the input screening problem have been proposed by several authors, and vary in the assumptions required, the sense in which importance is measured, and the number of model executions required for satisfactory performance. The four approaches reviewed in this paper are representatives of a large collection of ideas introduced as *uncertainty analysis* or *sensitivity analysis*. Our intent here is to point out the spectrum of compromises they offer between required assumptions and required evaluations.

2. ASSUMPTIONS, INPUT IMPORTANCE, AND MODEL RUNS

2.1. Linear Approximation

A time-honored and often useful assumption about a function of interest is that it is at least approximately linear in its arguments. This is such a strong assumption that it effectively boils the entire question of the behavior of m down to a single slope parameter for each input. There can be little question as to the definition of importance of any input in this case. The linearity assumption implies that $\partial y/\partial x_i$ takes the same value everywhere in Δ, which in turn fully defines any sense of how variable y is with respect to x_i. Complexity is not an issue here unless it is also defined so as to increase with the derivative.

Local sensitivity analysis often amounts, in practice, to definition of Δ to be small enough so that an assumption of approximate linearity is plausible. Downing et al. [3] are among the may authors who have described how first-difference approximations to partial derivatives can be derived from simple one-factor-at-a-time computer experiments. More recent practitioners of this approach sometimes use orthogonal 2-level fractional factorial designs of Resolution III or IV as the basis of such studies. Minimal designs supporting this kind of analysis generally contain from approximately k to $2k$ model runs, where k is the number of inputs.

Approaches requiring even fewer model evaluations may be developed if even stronger assumptions can be made. If it is reasonable to assume that most inputs have little or no effect on the output ("effect sparsity" in some literature) and/or the signs of each derivative can be assumed to be known, then group screening plans offer sequential strategies to identify important inputs using substantially fewer than k model evaluations. See the forthcoming book edited by Dean and Lewis [2] for a description of many of these strategies.

While this general approach is often useful and usually simple, one disadvantage is that there is little basis upon which to base an objective analysis of uncertainty. Since there is no formal basis for the statistical interpretation of residuals, quantities such as the t-statistics associated with each slope have only very limited heuristic value.

2.2. Input-Output Correlations

If approximate linearity is not a justifiable assumption, it still may be acceptable in some cases to assume that the slope of y with respect to x_i's, averaged over Δ, is an acceptable measure of input importance. This is probably most reasonable when an argument can be made that y is monotonic in the arguments of interest, and that the degree of nonlinearity in its behavior is limited. In these cases, an index such as

$$\int (y(\mathbf{x}) - \bar{y})(x_i - E(x_i)) f(\mathbf{x}) d\mathbf{x}, \quad \bar{y} = \int y(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

may be reasonable, where $y(\mathbf{x}) = m(\mathbf{x})$ and f is a probability density function.

Such integrals are easily estimated using a relatively modest Monte Carlo sampling of inputs, although most guidelines would suggest the need for more function evaluations than can be used when the strict linear approximation is used. The virtues of using Latin Hypercube sampling rather than unconstrained random sampling of inputs have been argued by McKay et al. [5] and Stine [13]. Iman and Conover [4] take this approach to evaluating the importance of inputs after transforming the output data to ranks.

The connection between k and the number of runs needed for effective Monte Carlo estimation of the integrals is not so clear as it is when a Linear Approximation is used. If more than a few inputs are important, accidental correlations between selected input values can be substantial unless the number of runs is not small compared to k. These problems may be moderated by using quasi-random sequences, e.g. [8], or algorithms such as Owen's [9] that control the degree of correlation between inputs.

2.3. Stochastic Continuity

Over the last 15 years or so, a number of papers have appeared in the statistics literature suggesting that the design and analysis of computer experiments might be based on regarding (1.) m as a realization of a spatial (i.e. Δ) stochastic process (frequentist), or (2.) the generalized uncertainty about m being expressed by such a process (Bayesian). See, e.g. Sacks et al. [10] and Currin et al. [1] for overviews of this approach. The most important practical issue in such approaches is the statement of a spatial covariance function, governing the "complexity" that may be expected in the behavior of the output as each input is varied. One popular functional form is:

$$Cov[y(\mathbf{x}), y(\mathbf{x}')] = \sigma^2 e^{-\sum_i \theta_i (x_i - x_i')^2}.$$

Given data from a computer experiment, likelihood or Bayes procedures may be used to estimate parameters such at the θ_i, and these used as importance indices. The sense of importance in this example function is, again, one of scale; the value of θ_i essentially defines distance in the x_i direction over which a given degree of activity would be expected in y.

Welch et al. [14] described an algorithm, for which the overall structure is much like that of stepwise regression, for identifying the inputs for which estimates of θ_i are largest, i.e. that are most important in this sense. In demonstrating the method, they evaluated two example functions each in $k = 20$ inputs using a Latin Hypercube sample of 50 runs. The methodology worked well in these exercises, but relatively few of the 20 inputs were actually important in each case; it might be reasonably expected that more runs would be needed if more of the inputs were active. The authors suggest that, following the identification of large correlation parameters, a sensible follow-up analysis would be examination of the fitted surface (mean of the conditional or posterior stochastic process) to examine the shape of m as a function of each apparently important input. However, reliable estimation of the response surface is likely to require more runs than reliable estimation of the covariance parameters.

One somewhat philosophical sticking point with (this version of) the Stochastic Continuity approach is that the indices of importance are parameters that do not *directly* describe properties of the function of interest! In the frequentist formulation of the problem, θ_i is a property of the (physically non-existant) process of which m is supposed to be a single realization. In the Bayesian model, θ_i is part of the characterization of a generalized uncertainty (or lack of understanding) of what the model might do under specified circumstances. With sufficient data (and I am not aware of a careful analysis of what this may mean in this application), this distinction may be less important practically than it is philosophically.

2.4. Conditional Variance

The approaches described to this point are predicated on assumptions of linearity, monotonicity, and continuity, respectively, in the model function. Even an assumption of continuity, however, is not always be warranted, and even when it is strictly warranted, the degree of complexity of y as a function of some x_i may make any attempt to explicitly model m difficult or impossible for practical purposes. In such cases it may be more natural or meaningful to define importance in purely statistical terms, e.g. the degree to which y may be expected to vary as x_i varies according to some (possibly arbitrary) probability distribution, completely disregarding any attempt to match a specific change in y to a specific change in x_i.

Sobol' [12], Saltelli et al. [11], and McKay [6] are among those who have proposed input sampling plans that support estimates of conditional moments of the distribution of y, where that distribution is propagated to the output from a specified distribution on the input vector. In particular, where each component of $\mathbf{x}$ is statistically independent of the others, these authors address estimation of

$$V_i[E_{(i)}[y(\mathbf{x})]] \quad \text{or "first-order variance"}$$
$$E_{(i)}[V_i[y(\mathbf{x})]] \quad \text{or "total variance"}$$

Here the subscript i means expectation or variance with respect to the marginal distribution of x_i, and subscript (i) implies the joint distribution of all inputs except x_i. No *functional* assumptions about m are involved here, but the nonparametric nature of this approach carries a practical requirement for a large number of model evaluations. Morris et al. [7] have recently identified other sampling plans based on Balanced Incomplete Block Designs that have some advantages for this type of analysis.

While this analysis has substantial appeal for the especially assumption-averse modeler, it also carries a philosophical difficulty (although not as fundamental as the one I described above for Stochastic Continuity methods). Here the objection is one of analysis efficiency. The indices of importance are estimated based entirely on the computed values of y, along with information about which runs share common randomly drawn values for each input. But the specific values of x_i are not used at all in the analysis; while they intuitively must carry some information of value in most practical situations, avoiding *all* assumptions about the y-to-$\mathbf{x}$ connection makes it difficult to apply this information.

3. COMPARISONS, CONCLUSIONS

The four general approaches outlined in Section 2 differ in (1.) the strength of assumptions that must be made about the model, (2.) the number of model evaluations that are required for practical purposes, and (3.) the sense in which importance is assessed for each input. Relatively strong assumptions leave relatively few degrees of freedom in defining importance, but require relatively few model evaluations for assessment. Relatively weak assumptions allow more subtle definitions of importance (or negatively, do not support the simplest interpretations), but require relatively many model evaluations.

Approach	Assumptions	Required Runs	Importance
Linear Approximation	most	least	derivative
Input-Output Correlations	↑	↓	averaged slope
Stochastic Continuity	↑	↓	complexity
Conditional Variance	least	most	variability

Variations on each of the approaches described here, and other fundamentally different approaches, have been proposed in the literature on computational science, applied mathematics, and statistics – the methods mentioned here are only an example of what has been found to be useful in many applications contexts. Future research might benefit from a broader inspection of how these methods differ, and how they might beneficially be combined to create new "points" along the assumption-data-interpretation spectrum.

REFERENCES

1. C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association* 86:953-963, 1991.

2. A.M. Dean and S.M. Lewis, eds. *Screening.* Springer-Verlag, New York, 2004.

3. D.J. Downing, R.H. Gardner, and F.O. Hoffman. An examination of response-surface methodologies for uncertainty analysis in assessment models. *Technometrics* 27:151-163, 1985.

4. R.L. Iman and W.J. Conover. Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Communications in Statistics - Theory and Methods* A9: 1749-1842, 1980.

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

6. M.D. McKay. Evaluation prediction uncertainty. Los Alamos National Laboratory Report NUREG/CR-6311, LA-12915-MS, 1995.

7. M.D. Morris, L.M. Moore, and M.D. McKay. Incomplete factorial sampling plans for evaluating the importance of computer model inputs. (Working paper, 2004)

8. H. Niederreiter. Quasi-Monte Carlo methods and pseudo-random numbers. *Bulletin of the American Mathematical Society* 84:957-1041, 1978.

9. A. Owen. Controlling correlations in Latin hypercube samples. *Journal of the American Statistical Association* 89:1517-1522, 1994.

10. J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science* 4:409-435, 1989.

11. A. Saltelli, T.H. Andrew, and T. Homma. Some new techniques in sensitivity analysis of computer model output. *Computational Statistics and Data Analysis* 15:211-238, 1993.

12. I.M. Sobol'. Sensitivity estimate for nonlinear mathematical models (in Russian). *Mat. Model.* 2, 1990.

13. M. Stein. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* 29:143-151, 1987.

14. W.J. Welch, R.J. Buck, J. Sacks, H.P. Wynn, T.J. Mitchell, and M.D. Morris. Screening, predicting, and computer experiments. *Technometrics* 34:15-25, 1992.