4

Variance-based Methods

How is variance decomposition related to sensitivity analysis? When is it worth using variance-based techniques instead of something else? Which measures can we obtain with variance-based methods? How are these measures calculated?

In this chapter we describe in more detail the variance-based methods that were introduced in Chapter 1. We first illustrate the settings that can be useful when dealing with modelling under conditions of uncertainty. We discuss the importance of a proper sensitivity test for a given setting. We sketch the historical background of variance-based methods, and discuss the properties of variance decompositions, from model independence to the capacity to evaluate the importance of groups of factors. Total effect indices are also introduced as a means of dealing synthetically with model complexity. We then illustrate two basic methods of computing the sensitivity indices, the Monte Carlo based design developed by Saltelli (2002) as well as the Random Balance Designs based on the Fourier Amplitude Sensitivity Test (FAST-RBD (Tarantola *et al.*, 2006)). Finally, we offer some examples.

4.1 DIFFERENT TESTS FOR DIFFERENT SETTINGS

It is common to find cases in the literature in which different sensitivity tests are applied to the same problem in a nonstructured fashion. This practice can yield a variety of results – e.g. in terms of ranking the factors in order of

Global Sensitivity Analysis. The Primer A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola © 2008 John Wiley & Sons, Ltd. ISBN: 978-0-470-05997-5

importance – with no guidance as to which we should believe or privilege. As discussed in Chapter 1, we suggest instead a careful consideration of (a) the output of interest and (b) the concept of 'importance', as it applies to the problem at hand. This would in general allow for the identification of the most appropriate setting for a given problem and, in turn, the sensitivity test to be applied. A list of possible settings (Saltelli *et al.*, 2004) is given here:

- The **Factor Prioritization** setting (FP) is used to identify a factor (or group of factors) which, when fixed to its true value, leads to the greatest reduction in the variance of the output. In other words, the identified factor (or group of factors) is that which accounts for most of the output variance. Therefore, this setting allows us (a) to detect and rank those factors which need to be better measured in order to reduce the output variance, as well as (b) to detect the factors that have a better chance of being estimated in a subsequent numerical or experimental estimation process. This latter point is particularly interesting as the analyst can identify the factors to be estimated before any estimation is made or measurements carried out. See Tarantola *et al.* (2000) for an example in the field of physics.
- The **Factor Fixing** setting (FF) is used to identify factors in the model which, left free to vary over their range of uncertainty, make no significant contribution to the variance of the output. The identified factors can then be fixed at any given value within their range of variation without affecting the output variance. This analysis can be performed on groups of factors, especially for large models, to identify noninfluential subsets of factors. Sometimes, factors are set up to represent alternative structures for model components (e.g. simple versus complex) and significant model simplifications can be often achieved when these factors are found to be noninfluential.
- The **Variance Cutting** setting (VC) is used for the reduction of the output variance to below a given tolerance. This may be desirable in risk or reliability analysis, where the analyst is interested in making sure, for example, that the uncertainty of the reliability of a given system component is below a given tolerance. In this setting the analyst wants to guarantee that the uncertainty is brought under a given value by acting on the smallest possible number of factors (see Saltelli and Tarantola, 2002).
- The **Factor Mapping** setting (FM) is used to study which values of the input factors lead to model realizations in a given range of the output space. For example, one may want to highlight model realizations falling above the 95th percentile because these correspond to risky conditions in an industrial plant or to a considerable financial loss (Campolongo *et al.*, 2007). In this setting one investigates which combination of factors leads to the realizations under analysis (Monte Carlo filtering, see Chapter 5).

The utility of variance-based sensitivity measures derives from their wide range of application. Of the four settings just recounted, the first three are susceptible of variance-based analysis.

4.2 WHY VARIANCE?

Most models live through their operational life as 'deterministic'. Each time they are 'interrogated' they are fed with a deterministic set of values for the input variables and the output – be it a scalar, a time series or a 2D map – is investigated for possible inferences. Sensitivity analysis for these models will generally entail changing one input at a time to test its effect on the output. In this book we profess a different philosophy of modelling, in which modellers are willing – and usually eager – to explore their model over different combinations of values for the uncertain inputs. Variance-based measures have proven useful in this framework. They study how the variance of the output depends on the uncertain input factors and can be decomposed accordingly.

But why study the variance? Could a sensitivity measure be built on the mean? To give an example, in risk analysis the model output may happen to be itself a mean,[1] and we might be interested in how the mean of the model output depends on its constituents. A legitimate question would then be how much each component of a system contributes to the risk that the system might fail. The answer could be that the risk depends 25% on component A, 15% on component B, and so on. Decomposing the absolute level of the risk into system components can be useful to help understanding which component is worth improving in order to reduce the level of risk in the system.[2]

Another measure encountered in risk analysis is the 'risk reduction worth', which measures the amount by which the risk associated with a system could be reduced if a model element were perfect, i.e. without risk of failure (see Borgonovo and Apostolakis, 2001).

The two examples just illustrated are based on a deterministic output – risk – and how the risk level can be modified by eliminating the uncertainty in the input. We are in principle against such an elimination of uncertainty; we prefer rather to retain the uncertain factors as an ingredient of sensitivity analysis. Even when we study how the mean of the output changes when

[1] For example, a risk may happen to be estimated as the product of the probability of a given outcome and the consequence of that outcome, averaged over a set of possible outcomes.

[2] Risk analysts use, for example, the Fussell–Vesely measure of importance in probabilistic safety assessments. This measure is defined as the fraction of risk that is contributed by the failure of a model element (Borgonovo and Apostolakis, 2001).

a factor is fixed – we denoted this in Chapter 1 as $E(Y|X_i = x_i^*)$ – we would then take the variance of this measure over all possible values x_i^*, i.e. $V(E(Y|X_i))$. Taking the mean of $E(Y|X_i = x_i^*)$ would have been of scant use – it would have led us back to the overall (unconditioned) mean. In other words, in a Monte Carlo framework variance emerges naturally if one wants to preserve the factors' uncertainty.

Returning to our example of risk analysis, note that 'risk' has been expressed as a crisp figure (e.g. a rate of failure, or the expected incidence of health effects), which may distract from the fact that risk is in itself an average. A practitioner might be interested in how the average is arrived at, in the form of the risk distribution tails and in details such as the topology of the low probability high-consequence regions and in the key assumptions shaping these regions. These issues are addressed by the methods presented in this book.

A key issue in sensitivity analysis is how to quantify the uncertainty of a model prediction – variance clearly being just one of the possible options. Depending on the problem at hand, we might be interested in the model prediction falling in the upper or lower 5th percentile of the distribution or in any particular interval of interest in the distribution, as in Monte Carlo filtering.

Methods have been developed which look at the entire distribution of the output and at how this is modified, on average, if a factor is fixed (Borgonovo, 2006).

We recommend using variance as a summary measure of uncertainty whenever the application allows it. This is in order to exploit the statistical properties of variance described in this chapter to investigate how factors contribute to the variance.

Interesting features of variance-based methods are:

- model independence: the sensitivity measure is model-free;
- capacity to capture the influence of the full range of variation of each input factor;
- appreciation of interaction effects among input factors;
- capacity to tackle groups of input factors: uncertain factors might pertain to different logical types, and it might be desirable to decompose the uncertainty according to these types.

The drawback of variance-based measures is their computational cost, as we shall discuss later in the chapter, and this is the reason why much recent research aims to find efficient numerical algorithms for their computation. Techniques for computation are offered both in this chapter and the next.

Sensitivity measures based on the decomposition of the variance of the model output are relatively recent in the literature. A brief summary of their development follows.

4.3 VARIANCE-BASED METHODS. A BRIEF HISTORY

Variance-based methods for sensitivity analysis were first employed by chemists in the early 1970s (Cukier *et al.*, 1973). Cukier and colleagues not only proposed conditional variances for a sensitivity analysis based on first-order effects, but were already aware of the need to treat higher-order terms and of the underlying variance decomposition theorems (Cukier *et al.*, 1978). Their method, known as FAST (Fourier Amplitude Sensitivity Test), although quite effective, enjoyed limited success among practitioners, not least because of the difficulty in encoding it. The method did not allow the computation of higher-order indices, although this was much later made possible by extensions developed by other investigators (see Saltelli *et al.*, 1999).

Also much later, Hora and Iman (1986) introduced the 'uncertainty importance' of a factor X_i,[3] defined as the expected reduction in the variance of the model output Y achieved by fixing X_i at a given value within its range of uncertainty:

$$I_i = \sqrt{\mathrm{Var}(Y) - E[\mathrm{Var}(Y|X_i)]}. \tag{4.1}$$

Later, the same authors (Iman and Hora, 1990) proposed a new statistic based on estimating the following quantity:

$$\frac{\mathrm{Var}_{X_i}[E(\log Y|X_i)]}{\mathrm{Var}[\log Y]}. \tag{4.2}$$

From Chapter 1 it is clear that this is the first-order variance term relative to $\log Y$. This measure has the advantage of robustness,[4] although it is not easy to convert results of sensitivity analysis pertaining to $\log Y$ back to Y. Transformations of Y for sensitivity analysis are presented in Chapter 5.

A visual inspection of sensitivity results has been suggested by Sacks *et al.* (1989). They proposed a decomposition of the response

$$Y = f(X_1, X_2, \ldots X_k) \tag{4.3}$$

into a set of functions of increasing dimensionality, whose plots are themselves used as measures of sensitivity (as will be discussed in detail in

[3] We refer to X_i as the ith element of $\mathbf{X}$, though the formulas presented in this chapter are appropriate also if X_i corresponds to a subset of model inputs.

[4] The range of variation of $\log Y$ can be much smaller than that of Y and hence its estimate can be obtained – *ceteris paribus* – at a lower sample size. For the same reason formulas similar to (4.2) were proposed on the rank of Y, e.g. replacing the values of Y with the integer corresponding to 1 for the highest Y value and with N (the size of the sample) for the lowest (Saltelli *et al.*, 1993).

Chapter 5). Although these authors do not use the variance, the functions they consider are the same as in Sobol's functional development.

The Russian mathematician I. M. Sobol' was inspired by the work of Cukier, and generalized it to provide a straightforward Monte Carlo-based implementation of the concept, capable of computing sensitivity measures for arbitrary groups of factors.

Given a square integrable function f over Ω^k, the k-dimensional unit hypercube,

$$\Omega^k = (X|0 \leq x_i \leq 1; i = 1, \ldots, k), \tag{4.4}$$

Sobol' considers an expansion of f into terms of increasing dimensions:

$$f = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + \ldots + f_{12\ldots k} \tag{4.5}$$

in which each individual term is also square integrable over the domain of existence and is a function only of the factors in its index, i.e. $f_i = f_i(X_i)$, $f_{ij} = f_{ij}(X_i, X_j)$ and so on. This decomposition is not a series decomposition, as it has a finite number of terms. Of the 2^k terms, one is constant (f_0), k are first-order functions (f_i),

$$\binom{k}{2}$$

are second-order functions (f_{ij}), and so on. This expansion, called high-dimensional model representation (HDMR), is not unique, meaning that, for a given model f, there could be infinite choices for its terms. Sobol' proved that, if each term in the expansion above has zero mean, i.e. $\int f(x_i)\mathrm{d}x_i = 0$, then all the terms of the decomposition are orthogonal in pairs, i.e. $\int f(x_i)f(x_j)\mathrm{d}x_i\mathrm{d}x_j = 0$. As a consequence, these terms can be univocally calculated using the conditional expectations of the model output Y.

In particular,

$$f_0 = E(Y) \tag{4.6}$$

$$f_i = E(Y|X_i) - E(Y) \tag{4.7}$$

$$f_{ij} = E(Y|X_i, X_j) - f_i - f_j - E(Y) \tag{4.8}$$

An analytic example of this decomposition is offered in Exercise 5.

As mentioned in Chapter 1 (see Conditional Variances – second path), the conditional expectation $E(Y|X_i)$ can be calculated empirically by cutting the X_i domain into slices and averaging the values of $(Y|X_i)$ within the same slice X_i. In this way, if the scatterplot has a pattern, the conditional expectation $E(Y|X_i)$ has a large variation across X_i values and the factor X_i is revealed to be important. Hence, the variance of the conditional expectation can be

considered as a summary measure of sensitivity. In fact, the variances of the terms in the decomposition above are the measures of importance being sought. In particular, $V(f_i(X_i))$ is $V[E(Y|X_i)]$; when we divide this by the unconditional variance $V(Y)$, we obtain the first-order sensitivity index. In short:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)}. \tag{4.9}$$

The first-order index represents the main effect contribution of each input factor to the variance of the output. The same quantity has been described by different investigators as an 'importance measure' (see Hora and Iman, 1986; Ishigami and Homma, 1996; Iman and Hora, 1990; Saltelli *et al.*, 1993; Homma and Saltelli, 1996), and as a 'correlation ratio' (see Krzykacz-Hausmann, 1990; McKay, 1996).

Sobol' also proposed a comparable definition of S_i (Sobol', 1996) which is based on the correlation between the model Y and the conditional expectation $E(Y|X_i)$:

$$S_i = Corr(Y, E(Y|X_i)). \tag{4.10}$$

4.4 INTERACTION EFFECTS

How can Sobol's variance decomposition help in investigating the existence of interaction effects? Two factors are said to interact when their effect on Y cannot be expressed as a sum of their single effects. Interactions may imply, for instance, that extreme values of the output Y are uniquely associated with particular combinations of model inputs, in a way that is not described by the first-order effects S_i just mentioned. Interactions represent important features of models, and are more difficult to detect than first-order effects. For example, by using regression analysis tools it is fairly easy to estimate first-order indices, but not interactions (remember the relationship $S_i = \beta_{X_i}^2$ discussed in Chapter 1 for linear models and orthogonal inputs, where β_{X_i} is the standardized regression coefficient for factor X_i).

A useful feature of decomposition (4.7 and 4.8) is that

$$V_i = V(f_i(X_i)) = V[E(Y|X_i)]$$

and

$$V_{ij} = V(f_{ij}(X_i, X_j)) = V(E(Y|X_i, X_j)) - V(E(Y|X_i)) - V(E(Y|X_j)).$$

In this equation, $V(E(Y|X_i, X_j))$ measures the joint effect of the pair (X_i, X_j) on Y, and, from now on we will denote the joint effect by V_{ij}^c. The term

$V(f_{ij})$ is the joint effect of X_i and X_j minus the first-order effects for the same factors. $V(f_{ij})$ is known as a second-order, or two-way, effect (Box *et al.*, 1978). Analogous formulas can be written for higher-order terms, enabling the analyst to quantify the higher-order interactions.

By condensing the notation of the variances, i.e. $V(f_i) = V_i$, $V(f_{ij}) = V_{ij}$, and so on, and by square integrating each term of the decomposition (4.5) over Ω^k, we can write the so-called ANOVA-HDMR decomposition:[5]

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \ldots + V_{12\ldots k}. \tag{4.11}$$

Dividing both sides of the equation by $V(Y)$, we obtain

$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij} + \sum_i \sum_{j>i} \sum_{l>j} S_{ijl} + \ldots + S_{123\ldots k} = 1 \tag{4.12}$$

which we already know from Chapter 1.

We recall that the number of these terms increases exponentially with the number of input factors.

Exercise 5, part 2, provides an example of the computation of partial variances and sensitivity indices for the same analytic case used for the functional decomposition in Section 4.3.

4.5 TOTAL EFFECTS

Total effects are a direct consequence of Sobol's variance decomposition approach and estimation procedure, although they were explicitly introduced and made computationally affordable by other investigators (see Homma and Saltelli, 1996; Saltelli, 2002).

The total effect index accounts for the total contribution to the output variation due to factor X_i, i.e its first-order effect plus all higher-order effects due to interactions.

For a three-factor model, for example, the total effect of X_1 is the sum of all the terms in Equation (4.12) where the factor X_1 is considered:

$$S_{T1} = S_1 + S_{12} + S_{13} + S_{123}. \tag{4.13}$$

---

[5] This is because of the orthogonality properties between any pair of terms in the expansion (see Exercise 5, part 3). Note that this variance decomposition holds only when the input factors X_i are independent (i.e. orthogonal). When the input factors are not independent of one another, the quantities V_i, V_{ij}^c, V_{ijl}^c retain their meaning but are no longer related to one another via (4.11).

In this example, the total index is composed of four terms. Total indices are useful in sensitivity analysis, as they give information on the nonadditive features of the model. As mentioned, for a purely additive model $\sum_{i=1}^{k} S_i = 1$, while for a given factor X_j a significant difference between S_{Tj} and S_j signals important interaction involving that factor. The total indices could be calculated in principle by computing all the terms in the decomposition (4.12), but there are as many as $2^k - 1$ of these. There are techniques that enable us to estimate total indices at the same cost of first-order indices (such as the Sobol' technique, see Homma and Saltelli (1996)), thus circumventing the so-called 'curse of dimensionality'. We customarily compute the set of all S_i plus the set of all S_{Ti} to obtain a fairly good description of the model sensitivities at a reasonable cost. We will see how to compute these indices in the next section.

The total effect measure provides the educated answer to the question: 'Which factor can be fixed anywhere over its range of variability without affecting the output?' The condition $S_{Ti} = 0$ is necessary and sufficient for X_i to be a noninfluential factor. If $S_{Ti} \cong 0$, then X_i can be fixed at any value within its range of uncertainty without appreciably affecting the value of the output variance $V(Y)$. The approximation error that is made when this model simplification is carried out depends on the value of S_{Ti} (see Sobol' *et al.*, 2007). Total indices are suitable for the factor fixing setting.

We recall from Chapter 1 that the unconditional variance can be decomposed into main effect and residual:

$$V(Y) = V(E(Y|X_i)) + E(V(Y|X_i)). \tag{4.14}$$

Another way to find the total index is to decompose the output variance $V(Y)$ again, in terms of main effect and residual, conditioning this time with respect to all the factors but one, i.e. $X_{\sim i}$:

$$V(Y) = V(E(Y|\mathbf{X}_{\sim i})) + E(V(Y|\mathbf{X}_{\sim i})). \tag{4.15}$$

The measure $V(Y) - V(E(Y|\mathbf{X}_{\sim i})) = E(V(Y|\mathbf{X}_{\sim i}))$ is the remaining variance of Y that would be left, on average, if we could determine the true values of $\mathbf{X}_{\sim i}$. The average is calculated over all possible combinations of $\mathbf{X}_{\sim i}$, since $\mathbf{X}_{\sim i}$ are uncertain factors and their 'true values' are unknown. Dividing by $V(Y)$ we obtain the total effect index for X_i:

$$S_{T_i} = \frac{E[V(Y|\mathbf{X}_{\sim i})]}{V(Y)} = 1 - \frac{V[E(Y|\mathbf{X}_{\sim i})]}{V(Y)}. \tag{4.16}$$

4.6 HOW TO COMPUTE THE SENSITIVITY INDICES

In this section we describe the Monte-Carlo based numerical procedure for computing the full set of first-order and total-effect indices for a model of k factors.

This procedure is the best available today for computing indices based purely on model evaluations. Additional procedures are described in the next chapter, based on emulators, i.e. on the ability to generate estimates of model output at untried points without rerunning the simulation model. The method offered here is attributable to Saltelli (2002) and represents an extension of the original approach provided by Sobol' (1990) and Homma and Saltelli (1996).

At first sight, it might seem that the computational strategy for the estimation of conditional variances such as $V(E(Y|X_i))$ and $V(E(Y|X_i, X_j))$ would be the cumbersome, brute-force computation of the multidimensional integrals in the space of the input factors. To obtain, for example, $V(E(Y|X_i))$, one would first use a set of Monte Carlo points to estimate the inner expectation for a fixed value of X_i, and then repeat the procedure many times for different X_i values to estimate the outer variance. To give an indication, if 1000 points were used to get a good estimate of the conditional mean $E(Y|X_i)$, and the procedure were repeated 1000 times to estimate the variance, then we would need 10^6 points just for one sensitivity index.

This is in fact not necessary, as the computation can be accelerated via existing short cuts. In the following, we describe the instrument proposed by Saltelli:

- Generate a (N, 2k) matrix of random numbers (k is the number of inputs) and define two matrices of data (A and B), each containing half of the sample (see 4.17 and 4.18). N is called a base sample; to give an order of magnitude, N can vary from a few hundreds to a few thousands. Sobol' recommends using sequences of quasi-random numbers (Sobol', 1967, 1976). The software to generate these sequences is freely available (**SIMLAB**, 2007).

$$A = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_i^{(1)} & \dots & x_k^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_i^{(2)} & \dots & x_k^{(2)} \\ \dots & \dots & \dots & \dots & & \\ x_1^{(N-1)} & x_2^{(N-1)} & \dots & x_i^{(N-1)} & \dots & x_k^{(N-1)} \\ x_1^{(N)} & x_2^{(N)} & \dots & x_i^{(N)} & \dots & x_k^{(N)} \end{bmatrix} \tag{4.17}$$

$$B = \begin{bmatrix} x_{k+1}^{(1)} & x_{k+2}^{(1)} & \dots & x_{k+i}^{(1)} & \dots & x_{2k}^{(1)} \\ x_{k+1}^{(2)} & x_{k+2}^{(2)} & \dots & x_{k+i}^{(2)} & \dots & x_{2k}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{k+1}^{(N-1)} & x_{k+2}^{(N-1)} & \dots & x_{k+i}^{(N-1)} & \dots & x_{2k}^{(N-1)} \\ x_{k+1}^{(N)} & x_{k+2}^{(N)} & \dots & x_{k+i}^{(N)} & \dots & x_{2k}^{(N)} \end{bmatrix}. \tag{4.18}$$

- Define a matrix C_i formed by all columns of B except the ith column, which is taken from A:

$$C_i = \begin{bmatrix} x_{k+1}^{(1)} & x_{k+2}^{(1)} & \dots & x_{i}^{(1)} & \dots & x_{2k}^{(1)} \\ x_{k+1}^{(2)} & x_{k+2}^{(2)} & \dots & x_{i}^{(2)} & \dots & x_{2k}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{k+1}^{(N-1)} & x_{k+2}^{(N-1)} & \dots & x_{i}^{(N-1)} & \dots & x_{2k}^{(N-1)} \\ x_{k+1}^{(N)} & x_{k+2}^{(N)} & \dots & x_{i}^{(N)} & \dots & x_{2k}^{(N)} \end{bmatrix}. \tag{4.19}$$

- Compute the model output for all the input values in the sample matrices A, B, and C_i, obtaining three vectors of model outputs of dimension $N \times 1$:

$$y_A = f(A) \qquad y_B = f(B) \qquad y_{C_i} = f(C_i). \tag{4.20}$$

We anticipate that these vectors are all we need to compute the first- and total-effect indices S_i and S_{Ti}, for a given factor X_i. Because there are k factors, the cost of this approach is $N + N$ runs of the model for matrices A, B, plus k times N to estimate k times the output vector corresponding to matrix C_i. The total cost is hence $N(k+2)$, much lower than the N^2 runs of the brute-force method.

Our recommended method estimates first-order sensitivity indices as follows:

$$S_i = \frac{V\left[E(Y|X_i)\right]}{V(Y)} = \frac{y_A \cdot y_{C_i} - f_0^2}{y_A \cdot y_A - f_0^2} = \frac{(1/N)\sum_{j=1}^{N} y_A^{(j)} y_{C_i}^{(j)} - f_0^2}{(1/N)\sum_{j=1}^{N} (y_A^{(j)})^2 - f_0^2} \tag{4.21}$$

where

$$f_0^2 = \left(\frac{1}{N}\sum_{j=1}^{N} y_A^{(j)}\right)^2 \tag{4.22}$$

is the mean, and the symbol $(\cdot)$ denotes the scalar product of two vectors.

Similarly, the method estimates total-effect indices as follows:

$$S_{T_i} = 1 - \frac{V\left[E(Y|\mathbf{X}_{\sim i})\right]}{V(Y)} = 1 - \frac{y_B \cdot y_{C_i} - f_0^2}{y_A \cdot y_A - f_0^2} = 1 - \frac{(1/N)\sum_{j=1}^{N} y_B^{(j)} y_{C_i}^{(j)} - f_0^2}{(1/N)\sum_{j=1}^{N} (y_A^{(j)})^2 - f_0^2}. \tag{4.23}$$

Why do these formulas work? We offer here a 'hand waving' explanation of (4.21). In the scalar product $y_A \cdot y_{C_i}$ values of Y computed from A are multiplied by values of Y for which all factors but X_i are resampled while the values of X_i remain fixed. If X_i is noninfluential, then high and low values of y_A and y_{C_i} are randomly associated. If X_i is influential, then high (or low) values of y_A will be preferentially multiplied by high (or low) values of y_{C_i} increasing the value of the resulting scalar product. We leave to the reader the task of understanding (4.23). (Hint: the scalar product $y_B \cdot y_{C_i}$ gives the first-order effect of non-X_i.)

Note that the accuracy of both f_0 and $V(Y)$ can be improved by using both y_A and y_B points rather than just y_A in Equations (4.21) and (4.23) (see Saltelli, 2002). This will improve the accuracy of the estimates for S_i and S_{Ti}, although the factors' ranking will remain unchanged. Error estimates for Equations (4.21) and (4.23) can be obtained by bootstrapping data points from vectors y_A, y_B and y_{C_i}. Alternatively, the error in the numerical estimates can be evaluated using the probable error associated with the crude Monte Carlo estimate.

The probable error is the error which will not be exceeded by 50 percent of the cases, and corresponds to 0.6745 [6] times the standard error.

For example, the probable error in $V_{X_i}[E(Y|X_i)]$ (that will not be exceeded by the error in the estimate with 50% probability) is:

$$P.E. = \frac{0.6745}{\sqrt{N}} \sqrt{\sum_{j=1}^{N} (y_A^{(j)} y_{C_i}^{(j)})^2 - \left(\sum_{j=1}^{N} y_A^{(j)} y_{C_i}^{(j)} \right)^2}.$$

Before using the sensitivity measures in our case studies, let us recall some of the properties of sensitivity indices that will prove useful in the interpretation of the results:

- Whatever the strength of the interactions in the model, S_i indicates by how much one could reduce, on average, the output variance if X_i could be fixed; hence, it is a measure of main effect.
- Whatever the interactions in the model, $S^c_{i_1,i_2,\ldots,i_s}$ indicates by how much the variance could be reduced, on average, if one could fix $X_{i_1}, X_{i_2}, \ldots, X_{i_s}$. We recall that 'c' denotes the joint effect.
- By definition, S_{Ti} is greater than S_i, or equal to S_i in the case that X_i is not involved in any interaction with other input factors. The difference

[6] The probability P_δ that a random sample from a normally distributed universe will have a mean m within a distance $|\delta|$ of the mean μ of the universe is $P_\delta = 2\Phi(|\delta|)$ where $\Phi(z)$ is the standard normal distribution function and δ is the observed value of $z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{N}}}$.
The value δ^* of δ such that $P_\delta = \frac{1}{2}$, is given by $\delta^* = \sqrt{2} \cdot erf^{-1}\left(\frac{1}{2}\right) = 0.6745$.

$S_{Ti} - S_i$ is a measure of how much X_i is involved in interactions with any other input factor.

- $S_{Ti} = 0$ implies that X_i is noninfluential and can be fixed anywhere in its distribution without affecting the variance of the output.
- The sum of all S_i is equal to 1 for additive models and less than 1 for nonadditive models. The difference $1 - \sum_i S_i$ is an indicator of the presence of interactions in the model.
- The sum of all $S_{T_i}s$ is always greater than 1. It is equal to 1 if the model is perfectly additive.

4.7 FAST AND RANDOM BALANCE DESIGNS

The classic Fourier Amplitude Sensitivity Test (FAST) method (Cukier *et al.*, 1978) is based on selecting N design points over a particular space-filling curve in the kth dimensional input space, built so as to explore each factor with a different (integer) frequency $(\omega_1, \omega_2, \ldots, \omega_k)$. A quite complex algorithm is used to set the frequencies such that they are free of interferences up to a given order M ($M = 6$ is usually considered sufficient). The computational model is run at each design point and the Fourier spectrum is calculated on the model output at specific frequencies $(\omega_i, 2\omega_i, \ldots M\omega_i)$ to estimate the sensitivity index of factor X_i. It is important that none of the higher harmonics of $\omega_1, \omega_2, \ldots, \omega_k$ interfere until order M, so that the Fourier spectrum at a given frequency corresponds uniquely to factor X_i. The design points are selected as follows:

$$X_i(s_j) = G_i(\sin \omega_i s_j), \quad \forall i = 1, 2, ..k, \quad \forall j = 1, 2, .., N \tag{4.24}$$

where X_i is the ith input factor, the functions G_i are chosen according to the desired pdf of X_i, s_j is the parametric variable varying in $(-\pi, \pi)$ which is sampled over its range using N points, and ω_i are the frequencies.

In Random Balance Designs (RBD) (Tarantola *et al.*, 2006) N points are selected over a curve in the input space using a frequency equal to 1 for each factor. The curve covers only a subset of the input space. Then independent random permutations are applied to the coordinates of the N points in order to generate the design points. The computational model is evaluated at each design point. Subsequently, the model outputs are reordered such that the design points are in increasing order with respect to factor X_i. The Fourier spectrum is calculated on the model output at the frequency 1 and at its higher harmonics (2, 3, 4, 5, 6) and yields the estimate of the sensitivity index of factor X_i. The same model outputs are reordered with respect to each other factor (and the Fourier spectra are calculated

accordingly) to obtain all the other sensitivity indices. The design points are chosen as follows:

$$X_i(s_{ij}) = G_i(\sin \omega s_{ij}), \quad \forall i = 1, 2, ..k, \quad \forall j = 1, 2, .., N \tag{4.25}$$

where $(s_{i1}, s_{i2}, \ldots, s_{iN})$ denotes the ith random permutation of the N points. Equation (4.25) provides a different random permutation for each factor X_i.

For RBD the model is evaluated N times over the sample of size N:

$$Y(s_j) = f(X_1(s_{1j}), X_2(s_{2j}), \ldots, X_k(s_{kj})) \quad \forall j = 1, 2, .., N. \tag{4.26}$$

The values of model output $Y(s_j)$, j = 1,..N are then reordered ($Y^R(s_j)$) such that the corresponding values of $X_i(s_{ij})$ are ranked in increasing order. The sensitivity of Y to X_i is quantified by the Fourier spectrum of the reordered model output:

$$F(\omega) = \left| \frac{1}{\pi} \sum_{j=1}^{N} Y^R(s_j) \exp(-\mathrm{i}\omega s_j) \right|^2 \tag{4.27}$$

evaluated at $\omega = 1$ and its higher harmonics (2, 3, 4, 5, 6). In the discrete case:

$$\widehat{V}_i = V[E(Y|X_i)] = \sum_{l=1}^{M} F(\omega)|_{\omega=l} = \sum_{l=1}^{M} R(l). \tag{4.28}$$

This is an estimate of the main effect V_i. The procedure is repeated for all factors, whereby the same set of model outputs is simply reordered according to $X_i(s_{ij})$ and (4.27) and (4.28) are used to estimate V_i, i= $2, \ldots, k$.

Here we provide the basic *Matlab*® code to compute a generic sensitivity index according to the RBD method:

```
s0=[-pi:2*pi/N:pi]';
s=s0(randperm(N))'; Performs a random permutation of the integers from 1 to N
x=.5+asin(sin(1*s))/pi; (see (4.25))
[dummy,index]=sort(s); orders the elements of s in ascending order
Y=model(x)
yr=y(index);
spectrum=(abs(fft(yr))).^2/N; fft is the fast Fourier transform
V1=2*sum(spectrum(2:M+1)); (see 4.28)
V=sum(spectrum(2:N));
S1=V1/V;
```

Random Balance Designs have a number of advantages with respect to FAST:

- the absence of a lower limit for the size N of the design points (FAST has the problem of aliasing, so a minimum sample size is required and this minimum size increases with the dimensionality k);
- the nonnecessity to have an algorithm to search for frequencies free of interferences;
- better accuracy in the estimates, which are not influenced by interferences;
- the possibility to select larger values of the order M without affecting the sample size N.
- contrary to the method of Saltelli, each model run contributes to the estimation of all the first-order indices.

The disadvantage of the RBD method is that it allows the computation of first-order terms only; we can use the sum of these to check if the model is additive. If the sum is noticeably smaller than 1 we must use another algorithm to compute interactions or total-effect terms. The main advantage of RBD is that it is relatively easy to implement, and the sample size N, being independent of the number of factors k, can lead to a considerable saving in computer time for expensive models.

4.8 PUTTING THE METHOD TO WORK: THE INFECTION DYNAMICS MODEL

Let us consider an infective process at its early stage, where I is the number of infected individuals at time t and S is the number of individuals susceptible to infection at time t.

We assume that the infection is propagated through some kind of contact between individuals who, especially at the early stage, do not take any precaution to avoid contagion.

It is reasonable to assume that the number of contacts per unit time is proportional to the number of individuals in each group (i.e. to $I \times S$) via a contact coefficient $k < 1$. Also, the number of infections is proportional to the number of contacts through an 'infection coefficient' ($\gamma < 1$), which is the likelihood that the infection is passed on during a given contact.

Depending on the dangerousness of the infection, the infected individuals will end up in either of two ways: by recovering or by dying. It is presumed that recovery and death rates (r and δ) are proportional to the number of infected individuals.

The number of susceptible individuals decreases with the number of infections, but can increase with new births b, proportional to S, or migration which happens at a constant rate m.

Two equations describe the dynamics of I and S, representing the model of the infection process:

$$\frac{dI}{dt} = \gamma kIS - rI - \delta I \tag{4.29}$$

$$\frac{dS}{dt} = -\gamma kIS + bS + m. \tag{4.30}$$

Let us investigate the evolution of the infection at its early stage $t \sim 0$, when we presuppose that the number of the susceptible individuals is much larger than that of the infected ($S(t) >> I(t)$), and that S is changing slowly ($S(t) \sim S_0 = const$).

Equation (4.29) becomes linear and homogeneous:

$$\frac{dI}{dt} = (\gamma kS_0 - r - \delta)I. \tag{4.31}$$

The solution is $I = I_0 \cdot exp(Y)$, where $Y = \gamma kS_0 - r - \delta$. If $Y > 0$ the infection spreads, while if $Y < 0$ the infection dies out.

Suppose that $S_0 = 1000$ (a small village), and that factors are distributed as follows:

- Infection coefficient $\gamma \sim U(0, 1)$. The infection is at an early stage, and no information is available about how it is acting.
- Contact coefficient $k \sim beta(2, 7)$. This distribution describes the probability of a person to come into contact with other individuals. In other words, the probability of meeting all the inhabitants of the village (and of meeting nobody) is low, while the probability of meeting an average number of persons is higher.
- Recovery rate $r \sim U(0, 1)$. We assume this to be uniform, as we do not know how it behaves at the beginning of the propagation.
- Death rate $\delta \sim U(0, 1)$, for the same reason as r.

Let us calculate the sensitivity indices for the four factors using the method described in Section 4.6. The total number of model runs is $N(k+2)$=7680 ($N = 1280$ and $k = 4$).

Let us analyse the sensitivity indices shown in Table 4.1. Negative signs are due to numerical errors in the estimates. Such negative values can often be encountered for the Saltelli method when the analytical sensitivity indices are close to zero (i.e. for unimportant factors). Increasing the sample size of the analysis reduces the probability of having negative estimates. FAST and RBD estimates are always positive, by construction.

The most influential factors are γ and k, while factors r and δ are noninfluential as their total indices are negligible. The infection is likely to

Table 4.1 First-order and total-effect sensitivity indices obtained with the method of Saltelli with $k \sim \text{beta}(2, 7)$

Factor	*First-order indices*	*Total-order indices*
Infection (γ)	0.49	0.69
Contact (k)	0.41	0.61
Recovery (r)	−0.00	−0.00
Death (δ)	−0.00	−0.00

spread proportionally to the number of contacts between people; unlike γ, which depends on the virus strength, k is a controllable factor. This means that if the number of contacts could be reduced, the variability in the output would also be reduced.

The sum of first-order effects is approximately 0.89, while the sum of the total indices is 1.29; as these two sums are both different from 1, there must be interactions among factors in the model. Moreover, given that both factors γ and k have total indices greater than their first orders, we conclude that they are taking part in interactions.

With the model outputs calculated at the 1280 points sampled above, we perform uncertainty analysis (see Figure 4.1, case 1). The picture shows that the infection propagates in almost all cases (in 99.7% of the cases the sign of the model predictions is positive).

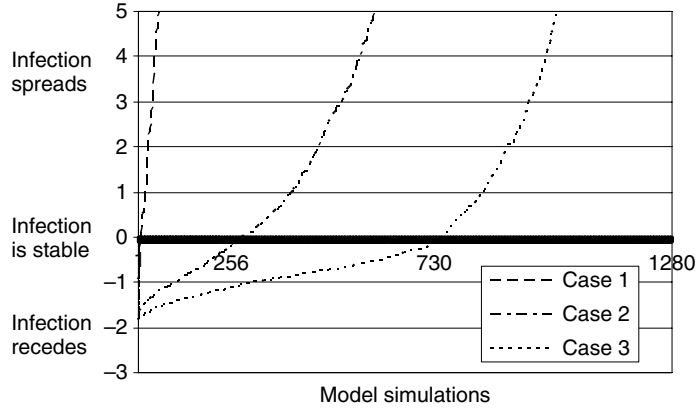

Figure 4.1 Uncertainty analysis for the three cases of the infection dynamics exercise. On the Y axis we plot the output variable of interest (if Y is positive the infection propagates, if it is negative the infection recedes and if it is zero – the X axis – the number of infected individuals is stable), and on the X axis the total number of model runs. Model outputs for each case are sorted in increasing order, so that each plot is a monotonic curve. As almost all model runs for the first case correspond to positive Y's, many of which can be of the order of a hundred, the Y axis is cut at +5 to visualize the plot around zero

Assume that, in consequence of this epidemic spread, some measures are taken in order to reduce the propagation of the disease. Following these new measures, which warn the inhabitants to avoid contacts, we assume that k is now distributed as $k \sim beta(0.5, 10)$ (see Figure 4.2) and we repeat the sensitivity analysis to see how the relative importance of the factors has been modified. The results are reported in Table 4.2.

We observe that, in this new configuration, factor k becomes more important in controlling the spread of the infection.

This causes us to wonder whether the restriction is adequate to reduce the propagation of the infection. By looking at the output of the model, we see that the infection recedes in 20% of cases, while in 80% of cases it propagates (see Figure 4.1, case 2): that is a significant improvement with respect to the initial situation, but stronger measures could still be taken.

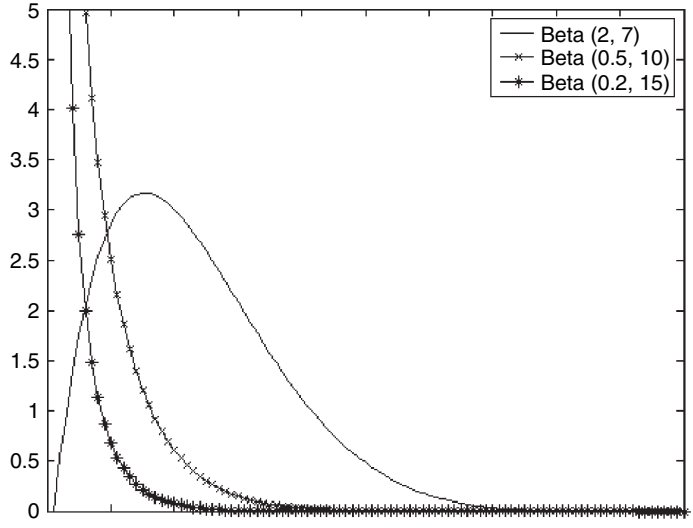

Figure 4.2 Beta distribution with different parameters

Table 4.2 First-order and total effect sensitivity indices obtained with the method of Saltelli with $k \sim \text{beta}(0.5, 10)$

Factor	*First-order indices*	*Total-order indices*
Infection (γ)	0.14	0.36
Contact (k)	0.76	0.98
Recovery (r)	−0.00	−0.00
Death (δ)	−0.00	−0.00

Assume that we implement additional restrictions on contact between persons, which means further squeezing the distribution of k to the left-hand side of its uncertainty range ($k \sim beta(0.2, 15)$, see Figure 4.2).

We observe in Table 4.3 that factor k becomes even more important, while factor γ has less influence in controlling the spread of the infection. The uncertainty analysis shows that in 57% of cases the infection recedes (see Figure 4.1, case 3).

The present example is quite academic, yet it shows how information obtained from sensitivity analysis (e.g. that the amount of contact between people is the most important factor in determining the spread of the disease) can help to inform decisions (e.g. designing measures to reduce people's contact in order to control the infection's propagation).

We test the RBD method described in Section 4.7 on the same case study selecting $N = 1280$, i.e. the same sample size that was employed for the method of Saltelli. In the RBD method each model run contributes to the estimation of all the sensitivity indices, while in the method of Saltelli it contributes to the estimation of one single first-order index (and its related total effect). In summary, RBD has better convergency properties than the method of Saltelli, in the sense that, for a given sample size, RBD estimates are more accurate (Tarantola *et al.*, 2006). We report the results for the three configurations of k in Table 4.4.

RBD produces indices for important factors (i.e. infection and contact) which are similar to those obtained with the method of Saltelli. Although the nonrelevant factors (recovery and death) are somewhat overestimated

Table 4.3 First-order and total-effect sensitivity indices obtained with the method of Saltelli with $k \sim \text{beta}(0.2, 15)$

Factor	*First-order indices*	*Total-order indices*
Infection (γ)	0.05	0.32
Contact (k)	0.77	1.05
Recovery (r)	−0.00	−0.00
Death (δ)	0.00	0.00

Table 4.4 RBD method first-order sensitivity indices

Factor	$k \sim beta(2, 7)$	$k \sim beta(0.5, 10)$	$k \sim beta(0.2, 15)$
Infection (γ)	0.43	0.13	0.05
Contact (k)	0.41	0.58	0.65
Recovery (r)	0.01	0.01	0.01
Death (δ)	0.01	0.01	0.01

by RBD, this is a minor problem, since they can anyhow be identified as noninfluential.

4.9 CAVEATS

Variance-based methods are powerful in quantifying the relative importance of input factors or groups. The main drawback of variance-based methods is the cost of the analysis, which, in the case of computationally intensive models, can become prohibitive even when using the approach described above.

With Saltelli's method, $N(k+2)$ runs for a full set of S_i and S_{Ti} require that for a model with 15 factors we need to execute the model at least 17 000 times, taking $N = 1000$. Using random balance designs with just N model executions we can compute the full set of S_i, but the S_{Ti} would remain unknown.

In terms of computational time, thousands or tens of thousands of model executions can be either trivial or unfeasible, depending on the model at hand. A viable alternative for computationally expensive models is the screening method discussed in Chapter 3. The elementary effect test is a good proxy for the total sensitivity indices.

If the model is both expensive to run and rich in factors we recommend using the elementary effect method to reduce the number of factors and then running a variance-based analysis on a reduced set of factors.

4.10 EXERCISES

Exercise 1

Let us consider the model $Y = \sum_{j=1}^{k} X_j$ where k=3, $X_j \sim U(\overline{x_j} - \sigma_j, \overline{x_j} + \sigma_j)$, $\overline{x_j} = 3^{j-1}$ and $\sigma_j = 0.5\overline{x_j}$.

Calculate the first-order sensitivity indices for the k factors.

First we calculate the expected value and the variance for the model.

$$
\begin{aligned}
Y &= X_1 + X_2 + X_3 \\
\overline{X_1} &= 1 \\
\overline{X_2} &= 3 \\
\overline{X_3} &= 9 \\
\sigma_1 &= 0.5 \\
\sigma_2 &= 1.5
\end{aligned}
$$

$$\begin{aligned}
\sigma_3 &= 4.5 \\
X_1 &\sim U(0.5, 1.5) \\
X_2 &\sim U(1.5, 4.5) \\
X_3 &\sim U(4.5, 13.5) \\
E(Y) &= E(X_1) + E(X_2) + E(X_3) = 1 + 3 + 9 = 13.
\end{aligned}$$

Next we compute the variance as

$$V(X_i) = E(X_i)^2 - E^2(X_i) = p(X_i) \int_a^b X_i^2 \mathrm{d}x_i - E^2(X_i). \tag{4.32}$$

So in our case we will have

$$\begin{aligned}
V(X_1) &= \int_{0.5}^{1.5} X_1^2 \mathrm{d}x_1 - 1 = \frac{1}{12} \\
V(X_2) &= \frac{1}{3} \int_{1.5}^{4.5} X_2^2 \mathrm{d}x_2 - 9 = \frac{3}{4} \\
V(X_3) &= \frac{1}{9} \int_{4.5}^{13.5} X_3^2 \mathrm{d}x_3 - 81 = \frac{27}{4} \\
V(Y) &= V(X_1) + V(X_2) + V(X_3) = \frac{1}{12} + \frac{3}{4} + \frac{27}{4} = \frac{91}{12}.
\end{aligned}$$

We compute the variance of the conditional expectation $V_{X_j}[E(Y|X_j)]$ and the expected residual variance $E_{X_j}[V(Y|X_j)]$.

$$\begin{aligned}
V_{X_1}[E(Y|X_1)] &= V(X_1) = \frac{1}{12} \\
V_{X_2}[E(Y|X_2)] &= V(X_2) = \frac{9}{12} \\
V_{X_3}[E(Y|X_3)] &= V(X_3) = \frac{81}{12} \\
E_{X_1}[V(Y|X_1)] &= V(Y) - V_{X_1}[E(Y|X_1)] = \frac{91}{12} - \frac{1}{12} = \frac{90}{12} \\
E_{X_2}[V(Y|X_2)] &= V(Y) - V_{X_2}[E(Y|X_2)] = \frac{91}{12} - \frac{9}{12} = \frac{82}{12} \\
E_{X_3}[V(Y|X_3)] &= V(Y) - V_{X_3}[E(Y|X_3)] = \frac{91}{12} - \frac{81}{12} = \frac{10}{12}.
\end{aligned}$$

Now we have all what we need to find the first-order indices:

$$S_1 = \frac{V_{X_1}[E(Y|X_1)]}{V(Y)} = \frac{1/12}{91/12} = \frac{1}{91}$$

$$S_2 = \frac{V_{X_2}[E(Y|X_2)]}{V(Y)} = \frac{9/12}{91/12} = \frac{9}{91}$$
$$S_3 = \frac{V_{X_3}[E(Y|X_3)]}{V(Y)} = \frac{81/12}{91/12} = \frac{81}{91}.$$

The model is additive, which means that there are no interactions among factors.

Exercise 2

Consider now the model $Y = X_1 + X_2$ where X_1, X_2 are normally distributed. We also know that $\overline{x_1} = 1, \overline{x_2} = 2$ and $\sigma_1 = 2, \sigma_2 = 3$.

Compute the expected value and the variance for model Y.

$$E(Y) = E(X_1) + E(X_2) = 1 + 2 = 3.$$

In this case, as factors are normally distributed, we can calculate the variance in an easier way:

$$V(Y) = V(X_1) + V(X_2) = \sigma_1^2 + \sigma_2^2 = 4 + 9 = 13. \tag{4.33}$$

Calculate the first-order sensitivity indices for the two factors:

$$S_1 = \frac{V_{X_1}[E(Y|X_1)]}{V(Y)} = \frac{4}{13}$$
$$S_2 = \frac{V_{X_2}[E(Y|X_2)]}{V(Y)} = \frac{9}{13}.$$

Also in this case the model is additive, without interactions among factors.

Exercise 3

Two input factors are normally distributed in the model

$$Y = X_1 \times X_2.$$

with parameters $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 3$ and $\sigma_2 = 2$.

Calculate the first- and second-order indices for the inputs and comment on the level of additivity of the model.

$$V(Y) = \mu_{X_1}^2 \sigma_{X_2}^2 + \mu_{X_2}^2 \sigma_{X_1}^2 + \sigma_{X_1}^2 \sigma_{X_2}^2 = 76.$$
$$S_{X_1} = \frac{\mu_{X_2}^2 \sigma_{X_1}^2}{V(Y)} = \frac{9}{19}$$

$$S_{X_2} = \frac{\mu_{X_1}^2 \sigma_{X_2}^2}{V(Y)} = \frac{1}{19}$$

$$S_{X_1,X_2} = \frac{\sigma_{X_1}^2 \sigma_{X_2}^2}{V(Y)} = \frac{9}{19}.$$

Factor X_1 is the most influential in determining the output variance (its first-order index is high). FactorX_2 has a low first-order index and is thus apparently less important.

Yet the interaction effect S_{X_1,X_2} is as high as the first-order effect of factor X_1. This means that the output variance is significantly driven by the two factors' interaction, even if factor X_2 appears to be noninfluential. This shows that ignoring interactions could lead to serious type II errors.[7]

Exercise 4

A model has eight input factors, but for computational cost's reasons we need to reduce the number of factors to five.

The model is

$$Y = \sum_{i=1}^{8} X_i$$

where X_i are normally distributed as follows:

$$X_1 \sim N(0, 1)$$
$$X_2 \sim N(2, 2)$$
$$X_3 \sim N(1, 3)$$
$$X_4 \sim N(1, 5)$$
$$X_5 \sim N(3, 1)$$
$$X_6 \sim N(4, 1)$$
$$X_7 \sim N(1, 2)$$
$$X_8 \sim N(5, 5)$$

1. Calculate the first-order sensitivity indices and identify the three least important factors, in order to exclude them from the model.
2. Recalculate the first orders for the remaining five factors and find out which are the most influential: if we decide to fix them at a given value in their range of variation, by what amount will the variance of the output be reduced?

[7] That is, ignoring the influence of an influential factor. This is typically the most serious, nonconservative error. On the other hand type I error means considering a noninfluential factor as influential.

1. We first calculate the output variance of the model:

$$V(Y) = \sum_{i=1}^{8} V(X_i) = 1 + 4 + 9 + 25 + 1 + 1 + 4 + 25 = 70.$$

Now it is easy to derive the first-order sensitivity indices for all input factors:

$$S_1 = \frac{V_{X_1}[E(Y|X_1)]}{V(Y)} = \frac{1}{70} = 0.01$$

$$S_2 = \frac{V_{X_2}[E(Y|X_2)]}{V(Y)} = \frac{4}{70} = 0.06$$

$$S_3 = \frac{V_{X_3}[E(Y|X_3)]}{V(Y)} = \frac{9}{70} = 0.13$$

$$S_4 = \frac{V_{X_4}[E(Y|X_4)]}{V(Y)} = \frac{25}{70} = 0.36$$

$$S_5 = \frac{V_{X_5}[E(Y|X_5)]}{V(Y)} = \frac{1}{70} = 0.01$$

$$S_6 = \frac{V_{X_6}[E(Y|X_6)]}{V(Y)} = \frac{1}{70} = 0.01$$

$$S_7 = \frac{V_{X_7}[E(Y|X_7)]}{V(Y)} = \frac{4}{70} = 0.06$$

$$S_8 = \frac{V_{X_8}[E(Y|X_8)]}{V(Y)} = \frac{25}{70} = 0.36.$$

The three least influential factors are X_1, X_5 and X_6, each one accounting for 1% of the output variance.

2. We discard the nonimportant factors from the model and we repeat the calculations with only the five remaining factors. We now have a new output variance:

$$V(Y) = \sum_{i=1}^{5} V(X_i) = 4 + 9 + 25 + 4 + 25 = 67,$$

and new elementary effects for the inputs:

$$S_1 = \frac{V_{X_1}[E(Y|X_1)]}{V(Y)} = \frac{4}{67} = 0.06$$

$$S_2 = \frac{V_{X_2}[E(Y|X_2)]}{V(Y)} = \frac{9}{67} = 0.14$$

$$S_3 = \frac{V_{X_3}[E(Y|X_3)]}{V(Y)} = \frac{25}{67} = 0.37$$
$$S_4 = \frac{V_{X_4}[E(Y|X_4)]}{V(Y)} = \frac{4}{67} = 0.06$$
$$S_5 = \frac{V_{X_5}[E(Y|X_5)]}{V(Y)} = \frac{25}{67} = 0.37.$$

We see that the two most influential factors are X_3 and X_5, each one determining 37% of the output variance.

If we decide to fix those two factors at a given value in their range of variation, we will have only three factors varying, i.e. factors X_1, X_2 and X_4. In such a situation, the model will have a lower variance:

$$V(Y) = \sum_{i=1}^{3} V(X_i) = 4 + 9 + 4 = 17.$$

We conclude that, in this example, by fixing the two most important factors the output variance decreases from 76 to 17, with a reduction of 75%.

Exercise 5

1. Calculate the expansion of f into terms of increasing dimensionality (4.5) for the function (Ishigami and Homma, 1996):

$$f(X_1, X_2, X_3) = \sin X_1 + a \sin^2 X_2 + bX_3^4 \sin X_1. \tag{4.34}$$

The input probability density functions are assumed as follows:[8]

$$p_i(X_i) = \frac{1}{2\pi},$$

when $-\pi \leq X_i \leq \pi$ and

$$p_i(X_i) = 0,$$

when $X_i < -\pi, X_i > \pi$ for $i = 1, 2, 3$.

[8] Note that this does not contradict the assumption that all factors are uniformly distributed within the unit hypercube Ω. It is always possible to map the hypercube to the desired distribution, and the sensitivity measure relative to the hypercube factors is identical to the measure for the transformed factors.

We calculate the decomposition of the function as (4.5) for $k = 3$:

$$f(X_1, X_2, X_3) = f_0 + f_1(X_1) + f_2(X_2) + f_3(X_3) + f_{12}(X_1, X_2) + f_{13}(X_1, X_3) + f_{23}(X_2, X_3) + f_{123}(X_1, X_2, X_3).$$

Thus

$$\begin{aligned} f_0 = E(Y) &= \int\int\int f(X_1, X_2, X_3)p(X_1)p(X_2)p(X_3)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3 \\ &= \frac{1}{(2\pi)^3}\int\int\int (\sin X_1 + a\sin^2 X_2 + bX_3^4 \sin X_1)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3 \\ &= \frac{1}{(2\pi)^3}\left[\int \sin X_1 \mathrm{d}x_1 + \int a\sin^2 X_2 \mathrm{d}x_2 + \int\int bX_3^4 \sin X_1 \mathrm{d}x_1\mathrm{d}x_3\right] \\ &= \ldots = \frac{a}{2}. \end{aligned}$$

So $f_0 = a/2$.

The $f_i(X_i)$ terms are easily obtained:

$$\begin{aligned} f_1(X_1) &= \int\int f(X_1, X_2, X_3)p(X_2)p(X_3)dx_2dx_3 - f_0 \\ &= \frac{1}{(2\pi)^2}\left[(2\pi)^2 \sin X_1 + a\int\int \sin^2 X_2 \mathrm{d}x_2\mathrm{d}x_3 \right. \\ &\quad \left. + b\sin X_1 \int\int X_3^4 \mathrm{d}x_2\mathrm{d}x_3\right] - \frac{a}{2} \\ &= \ldots = \frac{1}{(2\pi)^2}\left[(2\pi)^2 \sin X_1 + 2a\pi^2 + \frac{4}{5}b\sin X_1 \pi^6\right] - \frac{a}{2} \\ &= \sin X_1 + \frac{1}{5}b\pi^4 \sin X_1 = \left(1 + \frac{1}{5}b\pi^4\right)\sin X_1. \end{aligned}$$

$$\begin{aligned} f_2(X_2) &= \int\int f(X_1, X_2, X_3)p(X_1)p(X_3)\mathrm{d}x_1\mathrm{d}x_3 - f_0 \\ &= \frac{1}{(2\pi)^2}\left[\int\int \sin X_1 \mathrm{d}x_1\mathrm{d}x_3 + (2\pi)^2 a\sin^2 X_2 \right. \\ &\quad \left. + b\int\int X_3^4 \sin X_1 \mathrm{d}x_1\mathrm{d}x_3\right] - \frac{a}{2} \\ &= \ldots = a\sin^2 X_2 - \frac{a}{2}. \end{aligned}$$

$$\begin{aligned} f_3(X_3) &= \int\int f(X_1, X_2, X_3)p(X_1)p(X_2)\mathrm{d}x_1\mathrm{d}x_2 - f_0 \\ &= \frac{1}{(2\pi)^2}\left[\int\int \sin X_1 \mathrm{d}x_1\mathrm{d}x_2 + a\int\int \sin X_2^2 \mathrm{d}x_1\mathrm{d}x_2\right. \end{aligned}$$

$$+bX_3^4 \int\int \sin X_1 \mathrm{d}x_1 \mathrm{d}x_2 \Big] - \frac{a}{2}$$
$$= \ldots = 0.$$

The $f_{ij}(X_i, X_j)$ terms are computed as

$$f_{12}(X_1, X_2) = \int f(X_1, X_2, X_3) p(X_3) \mathrm{d}x_3 - f_1(X_1) - f_2(X_2) - f_0$$
$$= \sin X_1 + a \sin X_2^2 + b \sin X_1 \frac{1}{2\pi} \int X_3^4 \mathrm{d}x_3 - f_1(X_1) - f_2(X_2)$$
$$- f_0 = 0.$$
$$f_{13}(X_1, X_3) = \int f(X_1, X_2, X_3) p(X_2) \mathrm{d}x_2 - f_1(X_1) - f_3(X_3) - f_0$$
$$= \sin X_1 + a \frac{1}{2\pi} \int \sin X_2^2 \mathrm{d}x_2 + bX_3^4 \sin X_1 - f_1(X_1) - f_3(X_3) - f_0$$
$$= \ldots = \left(bX_3^4 - \frac{1}{5} b\pi^4 \right) \sin X_1.$$
$$f_{23}(X_2, X_3) = \int f(X_1, X_2, X_3) p(X_1) \mathrm{d}x_1 - f_2(X_2) - f_3(X_3) - f_0$$
$$= \frac{1}{2\pi} \int \sin X_1 \mathrm{d}x_1 + a \sin X_2^2 + bX_3^4 \sin X_1 dx_1$$
$$- f_2(X_2) - f_3(X_3) - f_0$$
$$= 0$$

f_{123} is obtained by difference and is equal to zero.

2. Calculate the variances of the terms for the function, according to Equation (4.11).

First we calculate the unconditional variance of the function:

$$V(f(X)) = \int [f(X_1, X_2, X_3) - E(f(X))]^2 p(X_1) p(X_2) p(X_3) \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3$$
$$= \frac{1}{(2\pi)^3} \int\int\int \left(\sin^2 X_1 + a^2 \sin^4 X_2 + b^2 X_3^8 \sin^2 X_1 \right.$$
$$+ 2a \sin X_1 \sin^2 X_2 + 2bX_3^4 \sin^2 X_1$$
$$+ 2abX_3^4 \sin X_1 \sin^2 X_2 + \frac{a}{4} \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3$$
$$\left. + a \sin X_1 - a^2 \sin^2 X_2 - abX_3^4 \sin X_1 \right)$$
$$= \frac{1}{(2\pi)^3} \left(\frac{1}{2} + \frac{3}{8} a^2 + \frac{b^2}{18} \pi^8 + \frac{a^2}{4} + \frac{b}{5} \pi^4 - \frac{a^2}{2} \right) (2\pi)^3$$
$$= \frac{1}{2} + \frac{a^2}{8} + \frac{b\pi^4}{5} + \frac{b^2 \pi^8}{18}.$$

We now calculate the variances of Equation (4.11), showing the passages for factor X_1:

$$V_1 = \int f_1^2(X_1)\mathrm{d}x_1 = \int \left(\sin X_1 + \frac{1}{5}b\pi^4 \sin X_1\right)^2 \mathrm{d}x_1$$
$$= \int \left[\sin X_1^2 + \frac{2}{5}b\pi^4 \sin X_1^2 + \frac{1}{25}b^2\pi^8 \sin X_1\right]\mathrm{d}x_1 =$$
$$= \ldots . = \frac{1}{2} + \frac{b\pi^4}{5} + \frac{b^2\pi^8}{50}.$$

The sensitivity index for factor X_1 can be calculated as

$$S_1 = \frac{V_1}{V} = \frac{1/2 + b\pi^4/5 + b^2\pi^8/50}{1/2 + a^2/8 + b\pi^4/5 + b^2\pi^8/18}$$

For the other factors we have

$$V_2 = \frac{a^2}{8}$$
$$V_3 = 0$$
$$V_{12} = 0$$
$$V_{13} = \frac{b^2\pi^4}{18} - \frac{b^2\pi^8}{50}$$
$$V_{23} = 0$$
$$V_{123} = 0$$

Again the fact that $V_{13} \neq 0$ even if $V_3 = 0$ is of particular interest, as it shows how an apparently noninfluential factor (i.e. with no main effect) may reveal itself to be influential through interacting with other parameters.

3. Show that the terms in the expansion of the function (4.34) are orthogonal.

Let us show, for example, that $f_1(X_1)$ is orthogonal to $f_2(X_2)$:

$$\int\int f_1(X_1)f_2(X_2)\mathrm{d}x_1\mathrm{d}x_2 = \left(1 + \frac{1}{5}b\pi^4\right)\int\int \sin x_1 \left(a\sin^2 x_2 - \frac{a}{2}\right)$$
$$= \left(1 + \frac{1}{5}b\pi^4\right)\int \sin x_1 \int \left(a\sin^2 x_2 - \frac{a}{2}\right)\mathrm{d}x_2,$$

which is equal to zero given that $\int \sin x_1 = 0$. The reader can verify, as a useful exercise, that the same holds for all other pairs of terms.