Sensitivity indices for imprecise probability distributions

Jim W. Hall[a]

[a]Department of Civil Engineering, University of Bristol, Queen's Building, University Walk, Bristol, BS8 1TR, UK
E-mail: jim.hall@bristol.ac.uk

Abstract: Conventional variance-based sensitivity indices are extended to deal with the case when information is available as closed convex sets of probability measures, a situation that exists when probability distributions are specified with interval-valued parameters. The generalization to closed convex sets of probability measures yields lower and upper sensitivity indices. An example demonstrates a numerical method for estimating these sensitivity indices.

Keywords: Variance-based sensitivity indices, coherent lower and upper probabilities

1. INTRODUCTION

The information input into computer models may be imprecise for several reasons. Imprecision is often a consequence of measurement processes, for example using digital sensors. Prior information is sometimes recorded in the literatures as intervals without any information about probability distributions [1]. Given only finite time, it is argued that it may be impossible to elicit precise probability distributions from experts [2]. Indeed experts may deliberately use imprecision to express their uncertainty.

The extension of probabilistic analysis to include imprecise information is now well established in the theory of imprecise probabilities [3], robust Bayesian analysis [4, 5] and fuzzy statistics [6]. In this paper we explore the notion of sensitivity within this framework. We confine ourselves to the theory of coherent lower and upper probabilities, which, whilst not the most general theory of imprecise probabilities, is sufficient to deal with the situation in which probability distributions are specified by interval-valued parameters.

2. COHERENT LOWER AND UPPER PROBABILITIES

Consider a probability density function $f(x, \mathbf{a})$, where $x \in \mathbb{R}$ and $\mathbf{a} = (a_1, a_2, \ldots, a_m)$, a vector of parameters of the probability density function. By definition

$$\Pr(A) = \int_A f(x, \mathbf{a}) dx, \forall A \subseteq \mathbb{R}. \tag{1}$$

If each parameter a_i in $\mathbf{a}$ is specified by a closed interval $[l_i, u_i]$ then $\mathbf{a}$ is constrained by an m-dimensional box Q, defining a closed set of probability measures that imply lower and upper probabilities, $P(\underline{A})$ and $P(\overline{A})$:

$$\Pr(\underline{A}) = \inf_{\mathbf{a} \in Q} \int_A f(x, \mathbf{a}) dx \tag{2}$$

Further author information: (Send correspondence to Jim W. Hall)
Jim W. Hall: Telephone: +44 117 928 9763

$$\Pr(\overline{A}) = \sup_{\mathbf{a}\in Q} \int_A f(x, \mathbf{a})dx. \tag{3}$$

$P(\underline{A})$ and $1 - P(\overline{A})$ will be located at the same point $\mathbf{a}$, so $P(\underline{A}) = 1 - P(\overline{A})$, meaning that $P(\underline{A})$ and $P(\overline{A})$ are coherent lower and upper probabilities [7].

The lower and upper expectations, $E(\underline{X})$ and $E(\overline{X})$, are given by

$$\underline{E}(X) = \inf_{\mathbf{a}\in Q} \int_{-\infty}^{\infty} xf(x, \mathbf{a})dx \tag{4}$$

$$\overline{E}(X) = \sup_{\mathbf{a}\in Q} \int_{-\infty}^{\infty} xf(x, \mathbf{a})dx. \tag{5}$$

The definitions in Equations 2 to 5 can be extended to the case when $f(\mathbf{x}, \mathbf{a})$ is a joint probability distribution on $\mathbb{R}^n$ and $\mathbf{x} = (x_1, \ldots, x_n)$.

2.1. Lower and upper variance

The standard definition of the variance $V(X)$ of a random variable X is

$$V(X) = E([X - E(X)]^2). \tag{6}$$

If $\mathcal{M}$ is a closed convex set of probability measures $P : X \to [0, 1]$, then the lower and upper variances $\underline{V}(X)$ and $\overline{V}(X)$ are given by:

$$\underline{V}(X) = \min_{P\in\mathcal{M}} V(X) \tag{7}$$

$$\overline{V}(X) = \max_{P\in\mathcal{M}} V(X). \tag{8}$$

2.2. Natural extension of imprecise probabilities

Let g be a function such that $y = g(\mathbf{x}) : \mathbf{x} = (x_1, \ldots, x_n)$, and let $B_y \subseteq \mathbb{R}^n$ containing all of the points $(x_1, \ldots, x_n)$ such that $g(\mathbf{x}) \in C : C \in \mathbb{R}$, then the lower and upper probabilities $\underline{P}(C)$ and $\overline{P}(C)$ are:

$$\underline{P}(C) = \inf_{\mathbf{a}\in Q} \int_{B_y} \cdots \int f(x_1, \ldots, x_n, \mathbf{a})dx_1 \ldots dx_n \tag{9}$$

and

$$\overline{P}(C) = \sup_{\mathbf{a}\in Q} \int_{B_y} \cdots \int f(x_1, \ldots, x_n, \mathbf{a})dx_1 \ldots dx_n. \tag{10}$$

3. VARIANCE-BASED SENSITIVITY ANALYSIS

Consider now the conventional probabilistic case in which the uncertainties in $x_1, \ldots, x_n$ are expressed as precise probability distributions, i.e. $x_1, \ldots, x_n$ and y are replaced by random variables $X_1, \ldots, X_n$ and Y respectively. In variance-based sensitivity analysis, the first order sensitivity indices S_i represents the fractional contribution of a given variable X_i to the variance in a given output variable Y [8]. In order to calculate the sensitivity indices the total variance V in the model output Y is apportioned to all the input factors X_i as [9]

$$V = \sum_i V_i + \sum_{i<j} V_{ij} + \sum_{i<j<k} V_{ijk} + \ldots + V_{12\ldots n} \tag{11}$$

where

$$V_i = V[E(Y|X_i = x_i^*)] \tag{12}$$

$$V_{ij} = V[E(Y|X_i = x_i^*, X_j = x_j^*)] - V_i - V_j \tag{13}$$

and so on. $V[E(Y|X_i = x_i^*)]$ is the Variance of the Conditional Expectation (VCE) and is the variance over all values of x_i^* in the expectation of Y given that X_i has a fixed value x_i^*. The first order (or 'main effect') sensitivity index S_i for variable X_i is:

$$S_i = V_i / V \tag{14}$$

and the 'total effect' sensitivity index is [10]

$$S_{Ti} = 1 - \frac{V[E(Y|X_{\sim i} = x_{\sim i}^*)]}{V(Y)} \tag{15}$$

where $X_{\sim i}$ denotes all of the variables other than X_i.

4. IMPRECISE SENSITIVITY INDICES

In the case when the uncertainty in the variables $X_1 \ldots X_n$ is described by a closed convex set $\mathcal{M}$ of probability measures P, the lower and upper variances introduced in Equations 7 and 8 above can be extended to lower and upper sensitivity indices, $\underline{S}_i$ and $\overline{S}_i$, $i = 1, \ldots, n$:

$$\underline{S}_i = \min_{P \in \mathcal{M}} S_i \tag{16}$$

and

$$\overline{S}_i = \max_{P \in \mathcal{M}} S_i \tag{17}$$

where

$$\sum_{i=1}^{n} \overline{S}_i \leq 1. \tag{18}$$

The additional constraint in Equation 18 means that the upper sensitivity indices $\overline{S}_i$, $i = 1, \ldots, n$ may not co-exist. Indeed there is a closed convex set $\mathcal{S}$ of sensitivity indices $\mathbf{S} \in \mathcal{S} : \mathbf{S} = \{S_1, \ldots S_n\}$ constrained such that $\forall S_i, i = 1, \ldots, n : \underline{S}_i \leq S_i \leq \overline{S}_i$ and $\sum_{i=1}^{n} \overline{S}_i \leq 1$.

4.1. Numerical method

Estimating the lower and upper sensitivity indices in Equations 16 and 17 is a problem of non-linear optimization. Each iteration j of the optimization involves estimating the precise sensitivity indices for some $P_j \in \mathcal{M}$, specified by a vector of parameters $\mathbf{a}_j = (a_1, \ldots, a_m)$. For each $\mathbf{a}_j$ the corresponding precise joint probability distribution $f(\mathbf{x}, \mathbf{a}_j)$ is randomly sampled d times, yielding a precise estimate of the variance [8]:

$$\hat{V}(Y_j) = \frac{1}{d} \sum_{k=1}^{d} g^2(\mathbf{x}_k, \mathbf{a}_j) - \hat{g}_{0,j}^2 \tag{19}$$

where

$$\hat{g}_{0,j} = \frac{1}{d} \sum_{k=1}^{d} g(\mathbf{x}_k, \mathbf{a}_j). \tag{20}$$

The Monte Carlo estimate $\hat{V}_i(Y_j)$ of the ith partial variance is given by

$$\hat{V}_i(Y_j) = \frac{1}{d} \sum_{k=1}^{d} g(\mathbf{x}_{\sim i,k}^{(1)}, \mathbf{x}_{i,k}^{(1)}, \mathbf{a}_j) g(\mathbf{x}_{\sim i,k}^{(2)}, \mathbf{x}_{i,k}^{(1)}, \mathbf{a}_j) - \hat{g}_{0,j}^2 \tag{21}$$

where

$$\mathbf{x}_{\sim i,k} = (x_{1,k}, x_{2,k}, \ldots, x_{i-1,k}, x_{i+1,k} \cdot \ldots, x_{n,k}). \tag{22}$$

The superscripts (1) and (2) in Equation 21 indicate that two sampling matrices are being used for $\mathbf{x}_k$. Both matrices have dimensions $d \times n$. In computing $\hat{V}_i(Y_j)$ the values of Y_j corresponding to $\mathbf{x}_k$ from matrix (1) are multiplied by the values of Y_j computed using a different matrix (2), but for the ith column, which is kept constant [8]. This resampling yields a precise estimate of the sensitivity indices $S_{i,j}$. The lower and upper variances are then given by

$$\underline{V}(Y) = \min_j(V(Y_j)) \tag{23}$$

$$\overline{V}(Y) = \max_j(V(Y_j)) \tag{24}$$

and the lower and upper sensitivity indices are given by

$$\underline{S_i}(Y) = \min_j(S_i(Y_j)) \tag{25}$$

$$\overline{S_i}(Y) = \max_j(S_i(Y_j)), i = 1, \ldots, n \tag{26}$$

where $S_i(Y_j) = V_i(Y_j)/V(Y_j)$.

5. APPLICATION

Oberkampf et al. [11] have proposed a series of Challenge Problems to compare and evaluate alternative theories of uncertainty. One of the Challenge Problems relates to a damped linear oscillator (a single degree of freedom mass-spring-damper system), whose steady-state magnification factor D_s is given by

$$D_s = \frac{k}{\sqrt{(k - m\omega^2)^2 + (c\omega)^2}} \tag{27}$$

where k is the spring constant, m is the mass of the oscillator, ω is the frequency of oscillation and c is the damping coefficient. In this Challenge Problem, the variables in Equation 27 were specified as follows:

m is given by a precise triangular probability distribution defined on the interval [10,12], with a median value 11.

k is given by an imprecise triangular probability distribution, specified by three imprecise parameters k_{min}, k_{mod} and k_{max}, whose values are contained in the closed intervals $k_{min} \in [90, 100]$, $k_{mod} \in [150, 160]$ and $k_{max} \in [90, 100]$.

c is given by a closed interval of possible values $c \in [5, 10]$. No probability distribution over this interval is specified or to be assumed.

ω is given by an imprecise triangular probability distribution, specified by three imprecise parameters ω_{min}, ω_{mod} and ω_{max}, whose values are contained in the closed intervals $\omega_{min} \in [2.0, 2.3]$, $\omega_{mod} \in [2.5, 2.7]$ and $\omega_{max} \in [3.0, 3.5]$.

In the Challenge Problem specification, the information concerning k and c was given by three independent sources. The problem of aggregation of evidence from multiple sources is beyond the scope of the present paper and is not addressed. The information is used from the first source only.

There are 6 interval-valued distribution parameters, k_{min}, k_{mod}, k_{max}, ω_{min}, ω_{mod}, ω_{max}, and one interval-valued variable, c, in the analysis. If the sensitivity indices S_i were a monotonic function of these imprecise quantities then it would only be necessary only to test the vertices of the 7 dimensional hypercube that contains all of the possible values of these quantities. There is, however, no reason to believe that S_i should be a monotonic function of these interval-valued quantities, so in order to find the imprecise sensitivity indices it was necessary to search the volume contained within these interval constraints. Besides testing each of the 2^7 vertices, the volume was searched by uniformly sampling the space with a total of 30000 samples. At each test point $\mathbf{a}_j = (k_{min,j}, k_{mod,j}, k_{max,j}, \omega_{min,j}, \omega_{mod,j}, \omega_{max,j}, c_j)$ (Equations 19 to 26) 50000 Monte Carlo samples were used in the sensitivity estimates.

The lower and upper upper probability distributions on D_s are shown in Figure 1. The lower and upper expectations were estimated as $\underline{E}(D_s) = 1.78$ and $\overline{E}(D_s) = 2.86$ and the lower and upper variances were estimated as $\underline{V}(D_s) = 0.09$ and $\overline{V}(D_s) = 1.57$. The imprecise sensitivity indices are listed in Table 1. Note the additional condition in Equation 18 means that the upper sensitivity indices cannot all coexist.

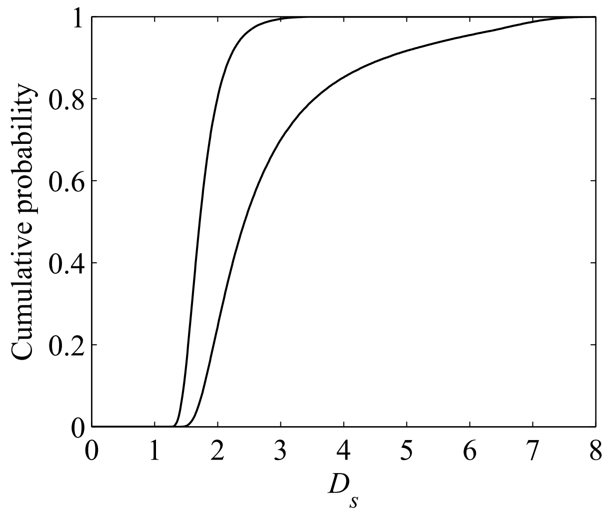

Figure 1. Lower and upper cumulative probability distributions of D_s

Table 1. Imprecise sensitivity indices

i	Variable	$\underline{S}_i$	$\overline{S}_i$
1	m	0.00	0.07
2	k	0.18	0.76
3	ω	0.19	0.70

6. CONCLUSIONS

Variance-based sensitivity indices provide an intuitive and practical expression of the contribution of model input variables to the variance in the model output [10, 12]. To date, variance-based sensitivity analysis have been restricted to the situation where uncertain information is presented as precise probability distributions, yielding precise sensitivity indices. In this paper this precise probabilistic case has been extended to the situation in which information appears as imprecise probability distributions or intervals, yielding interval-valued sensitivity indices for the (precise or imprecise) probabilistic variables. These imprecise indices complement the insights into the effects of imprecision and randomness provided by generalized uncertainty analysis [13]. A further challenge, which has not been addressed in this paper, is the problem of aggregation of imprecise and probabilistic information from multiple sources [14, 15]. Sensitivity analysis has further potential in this respect in highlighting the influence of different information sources.

The computational expense of calculating imprecise sensitivity indices is considerable. Furthermore, the advantage over Monte Carlo approaches of efficient methods for calculating variance-based sensitivity indices, such as FAST and Sobol' methods [8], is less clear than in the precise case. Monte Carlo methods can make use of function evaluations from previous steps in the optimization to find the lower and upper sensitivity indices, whereas the FAST and Sobol' methods would usually require a new sample at each optimization step. Whilst for the example addressed in this paper little computational advantage was to be gained by reusing previous function evaluations, clearly this will be desirable in many practical situations, so methods of this type are the subject of ongoing research.

ACKNOWLEDGMENTS

Dr Hall's research is supported by a Royal Academy of Engineering Post-Doctoral Research Fellowship.

REFERENCES

1. J.W. Hall, E. Rubio, and M.J. Anderson. Random sets of probability measures in slope hydrology and stability analysis. *ZAMM: Journal of Applied Mathematics and Mechanics*, in press, 2004.
2. J. Berger. The robust bayesian viewpoin (with discussion). In J. Kadane, editor, *Robustness of Bayesian Analyses*. North-Holland, Amsterdam, 1984.
3. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
4. P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
5. F.R. Hempel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, Chichester, 1986.
6. R. Viertl. *Statistical Methods for Non-Precise Data*. CRC Press, Boca Raton, Florida, 1996.
7. P. Walley. Towards a unified theory of imprecise probabilities. *Int. J. Approximate Reasoning*, 24(2-3):125–148, 2000.
8. K. Chan, S. Tarantola, A. Saltelli, and I.M. Sobol'. Variance-based methods. In A. Saltelli, K. Chan, and E.M. Scott, editors, *Sensitivity Analysis*, chapter 8, pages 167–198. Wiley, Chichester, 2000.
9. I.M. Sobol'. Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling Computational Experiment*, 1:407–414, 1993.
10. T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of model output. *Reliability Engineering and Systems Safety*, 52(1):1–17, 1996.
11. W.L. Oberkampf, J.C. Helton, C.A. Joslyn, S.F. Wojtkiewicz, and S. Ferson. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safe*, in press.
12. A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modelling. *Statistical Science*, 15(4):377–395, 2000.
13. S. Ferson. Probability bounds analysis is global sensitivity analysis. In *this volume*, 2004.
14. C. Genest and J.V. Zidek. Combining probability distributions: A critique and annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.
15. K. Sentz and S. Ferson. Combination of evidence in dempster-shafer theory. Technical Report SAND2002-0835, Sandia National Laboratories, Albuquerque, New Mexico, 2002.