Accelerated estimation of sensitivity indices using State Dependent Parameter models

Marco Ratto[a], Stefano Tarantola[a], Andrea Saltelli[a] and Peter C. Young[b]

[a]Institute for the Protection and Security of the Citizen
Joint Research Centre – European Commission
TP 361, Via E. Fermi 1, 21020 Ispra (VA), Italy;
[b]Centre for Research on Environmental Systems and Statistics
CRES/IENS,Lancaster University
Lancaster LA1 4YQ, U.K.
E-mail: marco.ratto@jrc.it, stefano.tarantola@jrc.it, andrea.saltelli@jrc.it, p.young@lancaster.ac.uk

Abstract: In this paper we use State Dependent Parameter (SDP) models (a non-parametric model estimation approach, based on recursive filtering and smoothing estimation) to estimate the main effect sensitivity indices of computational models. Especially when coupled with efficient sampling methods, such as the quasi-random LP-tau sequence, this method is extremely efficient, allowing for drastic reduction in the cost of the sensitivity analysis. Moreover, the method allows us also to estimate the first order terms of the High Dimensional Model Representation of the model under analysis, at no additional cost.

Keywords: State Dependent Parameter models, Variance based methods, High Dimensional Model Representation, Sensitivity analysis.

1. STATE OF THE ART

Consider the mathematical or computational model $Y = f(X_1, X_2, ..., X_k)$, where some of the input factors X_i are uncertain. For the non-correlated case, sensitivity indices are related to the Sobol' decomposition [1]

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + ... + V_{12...k} \tag{1}$$

where $V_i = V_{X_i}\left(E_{\mathbf{X}_{-i}}\left(Y|X_i\right)\right)$, $V_{ij} = V_{X_iX_j}\left(E_{\mathbf{X}_{-ij}}\left(Y|X_i, X_j\right)\right) - V_i - V_j$ and so on. This is also related to a decomposition of the function f itself into terms of increasing dimensionality (HDMR, [2-3]), i.e.

$$f(Y) = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + ... + f_{12...k} \tag{2}$$

where each term is a function only of the factors in its index, i.e. $f_i = f_i(X_i)$, $f_{ij} = f_{ij}\left(X_i, X_j\right)$ and so on. The various terms can be expressed as:

$$
\begin{aligned}
&f_0 = E(Y) \\
&f_i(X_i) = E_{\mathbf{X}_{-i}}(Y \mid X_i) - f_0 \\
&f_{ij}(X_i, X_j) = E_{\mathbf{X}_{-ij}}(Y \mid X_i, X_j) - f_i(X_i) - f_j(X_j) - f_0 \\
&\ldots
\end{aligned}
\tag{3}
$$

Variance-based sensitivity measures are based on the terms in the decomposition (1), normalised by the total unconditional variance $V(Y)$:

$$
\begin{aligned}
&S_i = V_i / V, \\
&S_{ij} = V_{ij} / V, \ldots \ S_{ij}^c = V_{X_i X_j}\left(E_{\mathbf{X}_{-ij}}\left(Y \middle| X_i, X_j\right)\right) / V \\
&\ldots
\end{aligned}
\tag{4}
$$

The S_i are called the main effects, the S_{ij} are the second order pure interaction terms, while the S_{ij}^c are the called the second order closed effects, giving the overall effect of a group of two factors, i.e. for orthogonal input factors $S_{ij}^c = S_i + S_j + S_{ij}$, and so on until the closed term of order *k* equal to 1. All indices are nicely scaled in [0, 1] and, as discussed in [4] and in the Keynote lecture by A. Saltelli at this conference [5], are related to rigorous settings, applicable to different contexts for SA. In particular, main effects are related to the setting "Prioritising Factors", i.e. to identify the factor which, if determined (i.e., fixed to its true, albeit unknown, value), would lead to the greatest reduction in the variance of the target output , and so on for the second most important factor etc., till all factors are ranked.

Let us then concentrate on the main effects S_i. The classical strategy for global sensitivity analysis methods is to directly estimate the V_i terms, without passing through the elementary functions f_i, and then to normalise by *V*. These methods (FAST, Extended FAST, correlation ratios, Sobol', etc, see [4-6] for reference) are conceived as black-box methods and do not try to use information present in the Monte Carlo sample, e.g. analysing scatter plots and trying some smoothing of the pattern, if any, between to model output and a given input. So, even if they are robust, unbiased and applicable to whatever non-linear and complex computational model, they do not make the best use of all the information contained in the Monte Carlo sample. This makes such methods computationally expensive, with a required number of model evaluations that is proportional to the number of factor *k*, e.g. of at least some thousands for a good approximation of the solution. This limits the application of variance based methods to not too complex computational models, which allow the required number of model evaluations to be carried out in a reasonable time. A lot of effort has been expended in recent years to reduce the cost of the analysis, either by improving the efficiency of the available methods (see e.g. [7]), or by exploring more efficient routes, such as the Bayesian approach presented by Oakley and O'Hagan [8]. In the latter case, Bayesian tools are used to exploit the information about the input-output mapping more efficiently than classical variance based methods, thus reducing drastically the computational cost of the analysis.

In this paper, we first estimate the f_i's, using recursive filtering and Fixed Interval Smoothing (FIS) algorithms to fit SDP models to the input-output mapping [9], then we

compute the variance of f_i to estimate the main effects. As in [8] this method allows us to estimate *both* f_i, and V_i (i.e. S_i), adding valuable information to the sensitivity analysis at a much smaller computational cost than classical methods. The convergence rate is of the same order of the Bayesian approach by Oakley and O'Hagan [8] and, at the same time, the method presented here is simple, since it is based on 'classical' recursive algorithms, such as the Kalman filter [10-11] and recursive FIS.

2. THE METHOD

The present methodology exploits signal processing and time series analysis tools, in particular an approach to non-stationary and nonlinear signal processing based on the identification and estimation of stochastic models with time variable (TVP) or state dependent (SDP) parameters. The works of P.C. Young [12-13] illustrate TVP/SDP algorithms and provide full references on the subject.

Often non-stationary and nonlinear systems can be approximated well by TVP (or piece-wise linear) models, the parameters of which can be estimated using recursive methods of estimation, where parameters are assumed to evolve in a simple stochastic manner (e.g. [12-14]). When instead the changes in the parameters are functions of the state or input variables (i.e. they actually constitute stochastic state variables), then the system is truly nonlinear and likely to exhibit severe nonlinear behaviour. Normally, this cannot be approximated in a simple TVP manner; in which case, the alternative and more powerful SDP modelling methods must be used.

In SDP time series modelling, the natural ordering of the data along the time coordinate is replaced by an ordering based on the ascending value of the state variables (or inputs), making the SDP model estimation similar to 'pattern recognition', i.e. to analysing scatter plots between a model input X_j and the output Y. In the SA framework, the analyst has a set of Monte Carlo simulations from which sensitivity indices and HDMR terms have to be estimated. Nothing impedes to consider such a set of Monte Carlo model evaluations as a time series and therefore to try to apply SDP modelling to estimate the first order terms in the decomposition of the computational model given in (2).

2.1. SDP models and HDMR

The general SDARX (State Dependent Auto-Regressive with eXogenous variables) specification for a dynamical system is:

$$Y_t = \mathbf{Z}_t^T \mathbf{p}_t + e_t \qquad e_t \sim N(0, \sigma^2) \tag{5}$$

where

$$\mathbf{Z}_t^T = [-Y_{t-1}, -Y_{t-2}, \ldots, -Y_{t-n}, \mathbf{X}_{t-\delta}^T, \mathbf{X}_{t-\delta-1}^T, \ldots, \mathbf{X}_{t-\delta-m}^T]$$

$$\mathbf{X}_t^T = [X_1, \ldots, X_k]$$

$$\mathbf{p}_t = [a_1(\mathbf{Z}_t), a_2(\mathbf{Z}_t), \ldots, a_n(\mathbf{Z}_t), \mathbf{b}_0(\mathbf{Z}_t), \mathbf{b}_1(\mathbf{Z}_t), \ldots, \mathbf{b}_m(\mathbf{Z}_t)]$$

and $a_i(\mathbf{Z}_t), i = 1,\ldots,n$ $\mathbf{b}_i(\mathbf{Z}_t), i = 0,\ldots,m$ are the state dependent parameters, which are assumed to be functions of the state vector $\mathbf{Z}_t$. We keep here the time series notation, with the index *t* spanning the set of Monte Carlo simulations, $t = 1,\ldots,N$.

In the SDP modelling framework the identification of the model structure itself is a critical issue. Observations of the input and output series are available and the analyst has to identify the dynamical model which best fits the observations, in the most parsimonious way (Data-Based Mechanistic modelling, [15]).

This involves finding which and how many lags and/or delays characterise the input state variables as well as if and how many autoregressive terms of the output have to be included. Moreover, the analyst has to make hypotheses on (i) which state variable are parameters dependent to and (ii) whether all parameters are state dependent or some of them are only time-dependent or simply constant.

In the present context, however, considerable simplifications can be achieved considering that model is deterministic and that, from (2), we *know* that, truncating all terms of order two and higher, the model can be written as:

$$Y_t - f_0 = f_1(X_{1,t}) + f_2(X_{2,t}) + \ldots + f_k(X_{k,t}) + e_t \quad e_t \sim N(0,\sigma^2) \tag{6}$$

where we assume that all terms of high order can be approximated by a Gaussian white noise with zero mean and variance σ^2, i.e. the truncated HDMR is seen as a stochastic non-linear system. This can be justified by a version of the central limit theorem [16], since the truncated terms can be seen as the sum of a large number of independent random variables with equal zero mean and arbitrary probability distribution.

Comparing the representation (6) to the SDP model definition (5) we can see that:

1. no autoregressive terms of the output variable are present in (6), i.e. *n*=0;
2. no lags or delays in the input variables are present in (6), i.e. *m*=0, δ=0;
3. items 1 and 2 imply that the state vector reduces to the vector of input variables, i.e. $\mathbf{Z}_t = \mathbf{X}_t$ and that the vector of time dependent parameters reduces to $\mathbf{p}_t = \mathbf{b}_0(\mathbf{X}_t)$;
4. each term of the sum (6) is a function of a single input variable, so each state dependent parameter b_{0i} depends only on the corresponding input variable X_i, i.e. $p_{i,t} = b_{0j}(\mathbf{X}_t) \equiv b_{0j}(X_{j,t})$. Without loss of generality, we can then re-write each term of (6) as $f_i(X_{j,t}) \equiv b_{0j}(X_{j,t})X_{j,t} = p_{j,t}X_{j,t}$.

So, the general, dynamic, time series specification (5), including lagged variables and delays, can be specialised to the HDMR of the computational model stopped to the first order (6) as follows:

$$Y_t - f_0 = p_{1,t}X_{1,t} + p_{2,t}X_{2,t} + \ldots + p_{k,t}X_{k,t} + e_t \qquad e_t = N(0,\sigma^2) \tag{7}$$

Estimating the state dependent parameters $p_{j,t}$ is hence equivalent to estimating the first order terms of the HDMR. The estimation of $p_{j,t}$ requires the following steps (see [12-13] for more details):

1. characterise the variability of $p_{j,t}$ in some stochastic manner: this is done using generalised random walk processes (GRW), specifically our choice is limited to the random walk (RW) and the integrated random walk (IRW) processes;

2. estimate via maximum likelihood (ML) the hyper-parameters (i.e. the variance of the innovations) of the RW/IRW processes of the state dependent parameters;

3. estimate the state dependent parameters and hence the first order terms of (2) applying recursive filtering and smoothing estimation, within an iterative backfitting procedure (The ML estimation in step 2. can also be iterated within the backfitting procedure).

With this procedure we estimate all the terms simultaneously, allowing us to use a single sample to estimate all indices. Moreover, the Monte Carlo sample is a standard one (pure random sample, Latin Hypercube, LP-tau, etc) and does not require a particular design, such as the classical variance based methods. This also allows it to be applied in the case of dependent inputs. We warn however that the convergence rate depends somehow on how the sample is generated. If quasi-random LP-tau random numbers are used, the convergence rate is very high, while using Latin Hypercube or pure random samples convergence is slower. This is clearly due to the more efficient exploration of the parameter space provided by the LP-tau quasi-random sequence.

2.2. The backfitting algorithm

The 'time scale' of the SDP model used for SA is just given by the sequence of the Monte Carlo evaluations of the computational model, so no 'logical' ordering can be expected in this sequence. In practice, the $\mathbf{p}_t$ values will continuously 'jump' in an extremely noisy way from one run to the subsequent in the Monte Carlo sample. So, it cannot be assumed that the simple GRW model is appropriate to describe such a variation over 'time'. However, it is possible to solve this problem if we sort the data in an ad-hoc manner. Specifically, if the ordering is chosen so that the SDP variations associated with the sorted series are smoother, it is more likely that a simple GRW process can be utilized to describe their evolution.

In our case, it is logical to assume that the most suitable ordering for each parameter $p_{j,t}$ should be done with respect to the corresponding input factor $X_{j,t}$. In this way, we can expect that the recursive estimation will be able to identify the pattern of Y_t vs. $X_{j,t}$. This also implies that each SDP needs a different sorting strategy, each with respect to its input factor. To solve this further problem, the backfitting procedure described in [13] can be exploited. Here, each parameter is estimated in turn, based on the modified dependent variable series obtained by subtracting all the other terms on the right hand side of (7) from Y_t. At each such backfitting iteration, the sorting can then be based on the single variable associated with the current SDP being estimated.

2.3. Specific issues for SA applications

2.3.1. Rescaling the inputs

Suppose that the input factor distribution of a factor X_j contains the zero. Suppose also that the corresponding f_j term assumes a non-zero value for $X_j = 0$. This poses a singularity problem, since, if $X_{j,t} = 0$ at a given sample point *t*, $p_{j,t}$ should assume an infinite value to yield a non-zero $f_{j,t}$. So, we propose that in the standard procedure for SA estimation, while leaving the output unmodified, all the input factors are transformed via a monotonic (invertible) operator into cumulative probabilities, obtaining a uniform distribution for all input factors in the [0, 1] interval. Then, to eliminate the zero, all values are then shifted to the interval [1, 2]. This allows the elimination of the singularity problems, while preserving the scale and shape of the model output.

Then, after estimation, the f_j^* terms estimated for the transformed variable X_j^*, can be plotted by putting in the axis of the abscissas the original values X_j, fully recovering the true f_j terms. In fact, if $P_j(X_j)$ is the cumulative distribution of factor X_j, at each sample point *t* we can write: $f_{j,t}^*(X_{j,t}^*) = f_{j,t}(P_j^{-1}(X_{j,t}^* - 1)) = f_{j,t}(X_{j,t})$.

2.3.2. The choice of the GRW model

A last methodological issue on the use of SDP models for sensitivity analysis of computational models concerns the choice of the GRW model, i.e. RW of IRW. Usually, sensitivity analysis tools, such as the software SIMLAB [17], aim to be 'black-box'. Whatever the model, inputs and outputs are fed to the SA tool to get the sensitivity indices. In principle, *a priori* one does not know whether RW or IRW model is more appropriate for each f_j term of the HDMR decomposition. In order to make this choice 'automatic', we propose the following preliminary step to the backfitting algorithm:

1. for each factor X_j perform the ML estimation of both RW and IRW models of the univariate model $Y_t = p_{j,t} X_{j,t} + e_t$;
2. for each estimated model, compute the R^2 measure or fit;
3. select the model with the highest R^2 in the subsequent SDP estimation.

3. APPLICATION

We have tested the method with several models, with up to 15 input factors. We first considered models with known analytical solutions for sensitivity indices and HDMR, such as the *g*-function of Sobol' [9]. Such tests allowed the convergence rate of the numeric to analytic solution to be measured. In general, a number of model runs of about 1,000 is sufficient for quite accurate estimates, with an absolute errors of about 0.01 on a scale [0, 1]. An extremely important improvement with respect to the classical estimation methods is that, in all the tests done, the computational cost was almost independent of the number of input factors. Clearly, it has to be expected that for a number of factors larger than 15-20, the convergence rate will start decreasing, but cases with such a large number of input factors are

usually preliminary passed through a screening method (e.g. the method of Morris or others) and only at a later stage the variance based/HDMR analysis is performed for fewer inputs. Convergence also becomes slower for non-monotonic models for which the parameters tend to be all rather equally important. Note that the latter condition is not usual, as most often factors follow a Pareto law, with few factors accounting for most of the importance. In all cases, the convergence is reached with a few thousand runs, with a considerable saving in computational effort with respect to the standard estimation methods (FAST, Sobol'). Even in the nastiest cases, convergence did not require more than 8000 runs.

Here, we show some significant results for the Level E model. Level E was used both as a benchmark of Monte Carlo computation [18] and as a benchmark for sensitivity analysis methods [19]. This test case has been extensively used by several authors; see [20] for a review. The model predicts the radiological dose to humans over geological time scales due to the underground migration of radionuclides from a nuclear waste disposal site. In a companion paper to this conference [21], readers can find another application of SDP modelling for the SA of a basin model to evaluate hydrocarbon exploration risk.

The Level E model has 12 input factors and is characterised by a strong non-linearity. Among the 12 parameters, X_4 (=$v^{(1)}$, water velocity in the first geosphere layer) and X_{12} (=W, stream flow rate) have the largest main effect over the simulated period. In Figures 1 and 2 we show the sensitivity indices versus time for these two parameters and compare the asymptotic values estimated with standard SA tools (Sobol' method), taking 1,000,000 runs, with the SDP estimation having total costs of 1024 and 8192. The samples for SDP model analysis were generated using LP-tau quasi-random sequences. We can see that already with only 1024 runs, which is a very small sample size for this kind of model, the absolute errors of the SDP estimates with respect to the asymptotic values is of the order of 0.01-0.02 in the sensitivity scale range of [0, 1]. With 1024 runs there is a critical point for W, where the drop of the sensitivity index at t=200,000 yr is shifted to the next time point t=300,000 yr. Increasing the total cost to 8192, results converge to the asymptotic values.

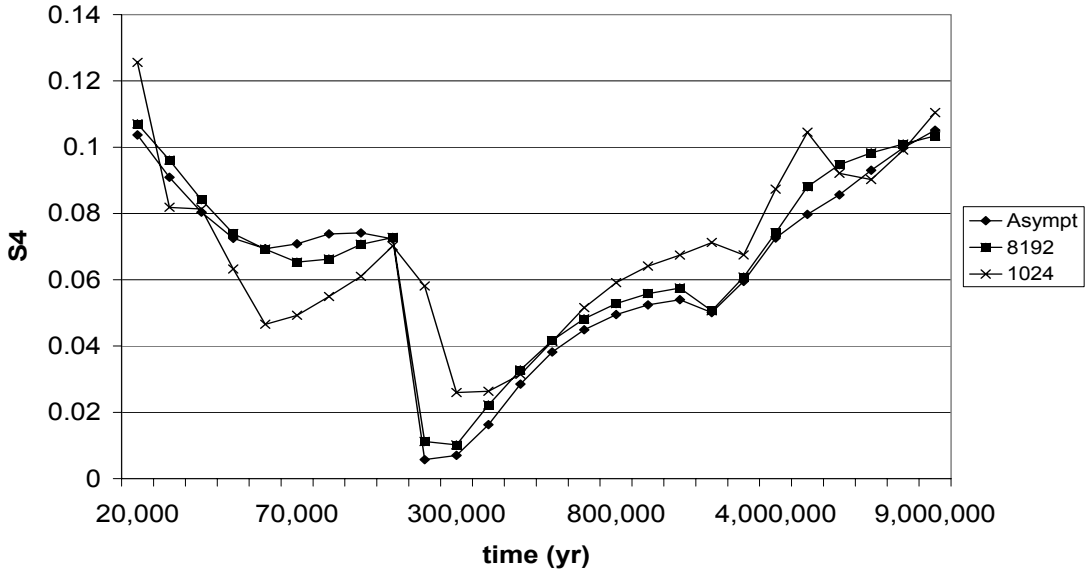

Figure 1. First order sensitivity index vs, time for parameter $v^{(1)}$ (X_4).

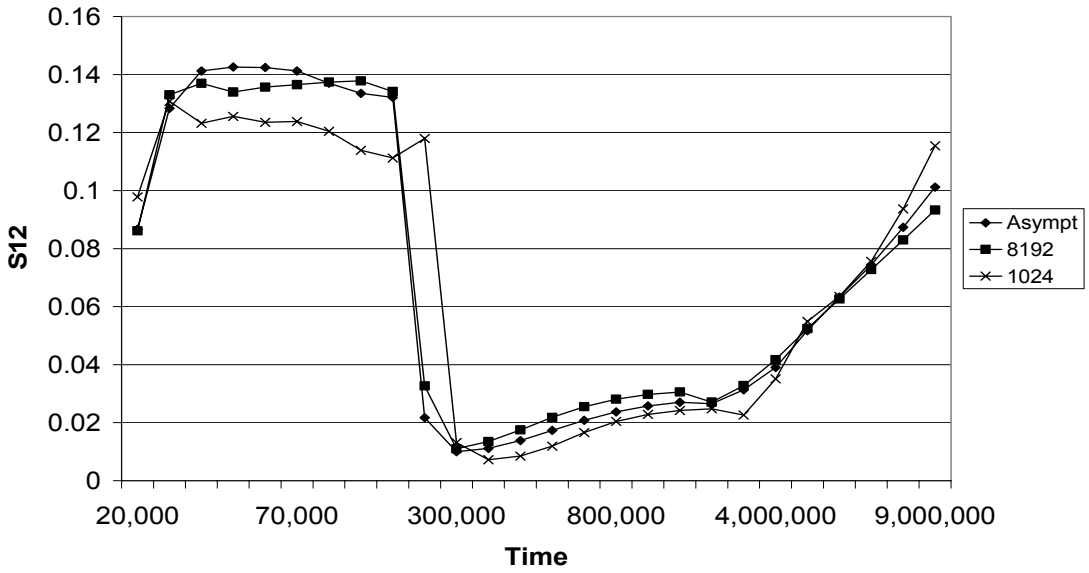

Figure 2. First order sensitivity index vs. time for parameter *W* (X12).

Comparing total costs, with the Sobol' technique we would need about 40,000 model runs to reach an accuracy comparable to the cheaper SDP model estimation of 1,024 model runs, i.e. the SDP modelling approach reduces the computational time by a factor 40 in this case. Conversely, 1024 runs for the Sobol' estimates are too few, with absolute errors that can reach 0.7-0.8, i.e. totally unreliable estimates.

In addition to sensitivity estimates, the SDP modelling approach also allows us to estimate the first terms in the high dimensional model representation decomposition. The plots of such functions for $v^{(1)}$ and *W* at the time *t*=100,000 are shown in Figure 3. The added value of the SDP modelling approach is evident by looking at the clear representation of the first order input-output mapping between $v^{(1)}$ and *W* and the output *Y* (the radiological dose). It is interesting to note in Figure 3 that the pattern estimated with 1024 runs for $f(X_4)$ slightly passes the zero axis for high values of X_4, while this is corrected increasing the number of runs to 8192. To better appreciate the 'pattern recognition' performed by the SDP estimation, in Figure 4 we also compare the scatter plots to the SDP estimates of $f(X_4)$ and $f(X_{12})$ (solid lines), for 8196 runs.

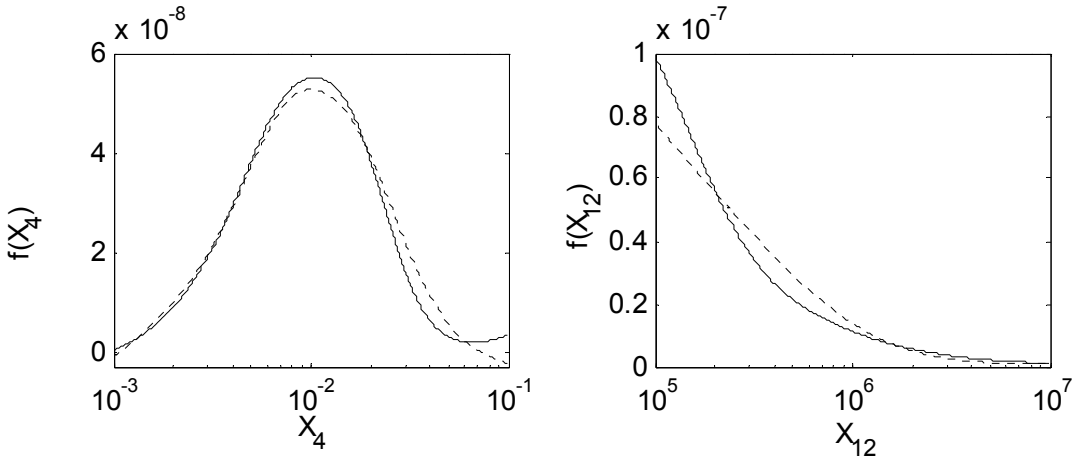

Figure 3. First order terms of the HDMR of the Level E model for $v(1)$ ($=X_4$) and W ($=X_{12}$) at t=100,000 yr. Solid lines are for the total cost of 8192 runs; dotted lines for the total cost of 1024 runs.

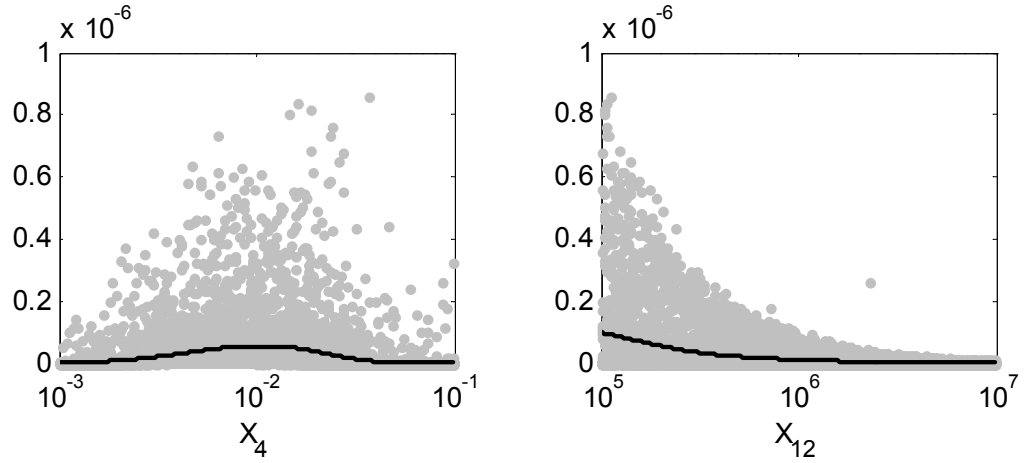

Figure 4. Scatter plots of Y and estimated first order terms of (2) for the Level E model for $v(1)$ ($=X_4$) and W ($=X_{12}$) at t=100,000 yr (8192 runs).

4. CONCLUSIONS

The use of SDP models is a powerful tool for a fast and accurate estimate of the first order terms of the HDMR and of the main effects sensitivity indices of computational models. All the estimates are performed with a unique sample, which can be any standard Monte Carlo sample. However, if efficient quasi-random number generators are used, such as the LP-tau sequence, the efficiency of the method is further enhanced, with a significantly faster convergence. We have tested the method with different models, always with extremely rapid convergence rates: 1,000 runs are in most cases sufficient for good estimates with models having up to 10-15 input factors. The dependence of the computational cost of the method to the number of input factors is very small: this is an extremely important improvement with respect to classical estimation methods. The convergence becomes slower in cases where the model is non-monotonic *and* the input factors share similar and relatively small levels of importance, i.e. they do not follow a Pareto law.

REFERENCES

1. I. M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 2: 112-118, 1990 (translated as: I. M. Sobol'. Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling & Computational Experiment*, 1: 407-414, 1993).
2. H. Rabitz, Ö. F. Aliş, J. Shorter, and K. Shim. Efficient input-output model representations. *Computer Physics Communications*, 117: 11-20, 1999.
3. H. Rabitz, and Ö. F. Aliş. Managing the Tyranny of Parameters in Mathematical Modelling of Physical Systems. In A. Saltelli, K. Chan and M. Scott, Eds, *Sensitivity Analysis*, 199-223. Wiley & Sons, Ltd, New York, 2000.
4. A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*, John Wiley & Sons publishers, Probability and Statistics series, to appear in March 2004.
5. A. Saltelli. Global Sensitivity Analysis: An Introduction. SAMO2004, Keynote lecture.
6. A. Saltelli, K. Chan, and M. Scott, Eds. *Sensitivity Analysis*. John Wiley & Sons, Ltd, New York, 2000.
7. A. Saltelli. Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, 145: 280-297, 2002.
8. J. Oakley, and A. O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Research Report No. 525/02* Department of Probability and Statistics, University of Sheffield, 2002. Submitted to *Journal of the Royal Statistical Society, Series B*.
9. Ratto M., Saltelli A., Young P., in preparation.
10. R. E. Kalman. A new approach to linear filtering and prediction problems. *ASME Trans., Journal Basic Eng.*, 82D: 35-45, 1960.
11. R. E. Kalman, and R. S. Bucy. New results in Linear filtering andpred iction theory. *ASME Trans., Journal Basic Eng.*, 83-D: 95-108, 1961.
12. P. C. Young. Nonstationary time series analysis and forecasting, *Progress in Environmental Science*, 1: 3-48, 1999.
13. P. C. Young. Stochastic, Dynamic Modelling and Signal Processing: Time Variable and State Dependent Parameter Estimation. Chapter from W. J. Fitzgerald et al. Eds, *Nonlinear and Nonstationary Signal Processing*, 74-114. Cambridge University Press, Cambridge, 2000.
14. P. C. Young. *Recursive Estimation and Time-Series Analysis*, Springer, 1984.
15. P.C. Young, S. D. Parkinson and M. Lees. Simplicity out of complexity: Occam's razor revisited. *Journal of Applied Statistics*, 23: 165-210, 1996.
16. W. Feller. The fundamental limit theorems in probability. *Bull. Amer. Math.Soc*., 51:800–832, 1945.
17. SIMLAB, Version 2.2, Simulation environment for Uncertainty and Sensitivity Analysis, http://www.jrc.cec.eu.int/uasa/prj-sa-soft.asp.
18. OECD/NEA PSAC User group. *PSACOIN Level E intercomparison.* B. W. Goodwin, J. M. Laurens, J. E. Sinclair, D.A. Galson, and E. Sartori., Eds. Nuclear Energy Agency, Organisation for Economic Cooperation and Development, Paris, 1989.
19. OECD/NEA PSAG User group. *PSACOIN Level S intercomparison*. A. Alonso, P. Robinson, E. J. Bonano and D. A. Galson, Eds. Nuclear Energy Agency, Organisation for Economic Cooperation and Development, Paris, 1993.
20. A. Saltelli, and S. Tarantola. On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. *Journal of American Statistical Association*, 97: 702-709, 2002.
21. L. P. Ruffo, L. Bazzana, A. Corradi, A. Saltelli. S. Tarantola, M. Ratto, Exploration Risk Evaluation Through UASA Techniques, SAMO2004, oral presentation SAMO-04-20.