

## Accelerated estimation of sensitivity indices using State Dependent Parameter models

M. Ratto, S. Tarantola, A. Saltelli  
Institute for the Protection and Security of the Citizen  
Joint Research Centre – European Commission  
TP 361, Via E. Fermi 1, 21020 Ispra (VA), Italy  
e-mail: [marco.ratto@jrc.it](mailto:marco.ratto@jrc.it)  
phone: +39 0332 785639, fax: +39 0332 785733

P. Young  
Centre for Research on Environmental Systems and Statistics  
CRES/IEENS, Lancaster University  
Lancaster LA1 4YQ, U.K.

### 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.

### State-of-the-art

Consider the mathematical or computational model  $Y = f(X_1, X_2, \dots, X_k)$ , where some of the input factors  $X_i$  are uncertain. For the non-correlated case, sensitivity indices are related to the decomposition (Sobol', 1990-1993)

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

where  $V_i = V(E(Y|X_i))$ ,  $V_{ij} = V(E(Y|X_i, X_j)) - V_i - V_j$  and so on. Sobol' decomposition is based on a decomposition of the function  $f$  itself into terms of increasing dimensionality (HDMR, H. Rabitz et al. 1999, 2000), i.e.,

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

where each term is a function only of the factors in its index, i.e.

$f_i = f_i(X_i)$ ,  $f_{ij} = f_{ij}(X_i, X_j)$  and so on. The various terms can be expressed as:

$$f_0 = E(Y)$$

$$f_i(X_i) = E(Y | X_i) - f_0 \tag{3}$$

$$f_{ij}(X_i, X_j) = E(Y | X_i, X_j) - f_i(X_i) - f_j(X_j) - f_0$$

...

Let us now concentrate on the main effects  $V_i$ , which provide a very important measure of sensitivity. The classical strategy for global sensitivity analysis methods is to estimate the  $V_i$  terms directly, without passing through the elementary functions  $f_i$ . These methods (FAST, Extended FAST, correlation ratios, Sobol', etc, see the Handbook edited by Saltelli et al., 2000, for reference) are conceived as black-box methods and do not try to use information present in patterns, 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 of some thousands for a good convergence to 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. Saltelli 2002), or by exploring new routes, such as the Bayesian approach presented by Oakley and O'Hagan (2002). 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 significantly the computational cost of the analysis.

In this paper, we first estimate  $f_i$ , and then compute the variance of  $f_i$ , using recursive filtering and Fixed Interval Smoothing (FIS) algorithms to fit SDP models to the input-output mapping (Ratto et al., 2003). This method allows us to estimate *both*  $f_i$ , and  $V_i$ , 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 but, at the same time, the method presented here is simpler, since it is based on 'classical' recursive algorithms, such as the Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) and recursive FIS.

## The method

The present methodology exploits signal processing and time series analysis tools, in particular an approach to nonstationary and nonlinear signal processing based on the identification and estimation of stochastic models with time variable (TVP) or state dependent (SDP) parameters. The reader should consult Young (1999, 2000), who develops these TVP/SDP algorithms and provides full references on the subject.

Although such nonstationary and nonlinear systems exhibit nonlinear behaviour, this can often be approximated well by TVP (or piece-wise linear) models, the parameters of which can be estimated using recursive methods of estimation in which the parameters are assumed to evolve in a simple stochastic manner (e.g. Young, 1984, 1999). On the other hand, if 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, recourse must be made to the alternative, and more powerful SDP modelling methods.

There is no reason why we should not consider the set of Monte Carlo model evaluations as a time series, and the SDP modelling method can be applied to these data in order to estimate the first order terms of the model decomposition. 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'. This is equivalent to analysing scatter plots between a model input  $X_j$  and the output  $Y$  and so allows SDP modelling to be used as a method of estimating the first order terms in the decomposition of the computational model given in (2).

A SDP model definition suitable for estimating the first order terms of the model decomposition can take the form,

$$Y^{(i)} - f_0 = b_1(X_1^{(i)})X_1^{(i)} + b_2(X_2^{(i)})X_2^{(i)} + \dots + b_k(X_k^{(i)})X_k^{(i)} + e_i \quad e_i = N(0, \sigma^2) \quad (4)$$

where  $b_j(X_j^{(i)})$ ,  $j = 1, \dots, k$ , are the state dependent parameters. It can be seen that each term of the sum (4) is a function of only one model input, i.e. the representation (4) is equivalent to the model decomposition (2) limited to the first order terms. It is further assumed that all the remaining terms behave like a white noise, i.e. the model is seen as a stochastic non-linear system.

Now the index  $i$  indicates the  $i$ -th model evaluation of the Monte Carlo sample, i.e.  $i=1, \dots, N$ . To make explicit the link with signal processing, the standard signal application field of SDP modelling, the index  $i$  should be replaced by  $t$ , indicating the time co-ordinate (so,  $t=1, \dots, N$ ).

The SDP model estimates 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. However, the convergence rate depends 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 significantly slower. This is clearly due to the more efficient exploration of the parameter space provided by the LP-tau quasi-random sequence.

### **Application**

We have tested the method with different models, always with extremely rapid convergence rates. Here, we show some significant results for the Level E model. Level E was used both as a benchmark of Monte Carlo computation (OECD 1989) and as a benchmark for sensitivity analysis methods (Level S, OECD 1993). This test case has been extensively used by several authors, see Saltelli and Tarantola (2002) 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.

The 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 Figure 1 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 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 estimates is of the order of 0.01-0.02 in a sensitivity scale range of [0, 1]. There is only a critical point for  $W$ , where the drop of the sensitivity index at  $t=10$  is shifted to  $t=11$ . Moreover, on increasing the total cost to 8192, convergence is already attained. 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 model decomposition. The plots of such functions for  $v^{(1)}$  and  $W$  at the time  $t=10$  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).

## Conclusions

The use of SDP models is a powerful tool for a fast and accurate estimate of main effect 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. The method allows us also to estimate the first order terms of the HDMR of the model under analysis at no additional cost.

## References

- Kalman, R.E. (1960) A new approach to linear filtering and prediction problems, *ASME Trans., Journal Basic Eng.*, **82D** 35-45.
- Kalman, R. E. and Bucy, R. S. (1961) New results in Linear filtering and prediction theory, *ASME Trans., Journal Basic Eng.* **83-D**, 95-108.
- Oakley, J. and O'Hagan, A. (2002). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Research Report No. 525/02* Department of Probability and Statistics, University of Sheffield. Submitted to *Journal of the Royal Statistical Society, Series B*
- OECD/NEA PSAC User group (1989), "PSACOIN Level E intercomparison. An international code intercomparison exercise on a hypothetical safety assessment case study for radioactive waste disposal systems", prepared by B. W. Goodwin, J. M. Laurens, J. E. Sinclair, D.A. Galson, and E. Sartori., Paris, OECD - NEA.
- OECD/NEA PSAG User group (1993), "PSACOIN Level S intercomparison. an international code intercomparison exercise on a hypothetical safety assessment case study for radioactive waste disposal systems", Paris, OECD - NEA.
- Rabitz, H., Ö. F. Aliş, J. Shorter, and K. Shim (1999), "Efficient input-output model representations", *Computer Physics Communications*, 117: 11-20.
- Rabitz, H., and Ö. F. Aliş, 2000, "Managing the Tyranny of Parameters in Mathematical Modelling of Physical Systems", in *Sensitivity Analysis*, A. Saltelli, K. Chan and M. Scott Eds., 199-223.
- Ratto M., Saltelli A., Young P., (2003), in preparation.
- Saltelli, A. (2002), Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, **145**, 280-297.
- Saltelli A., K. Chan K., and M. Scott, Eds, (2000), "*Sensitivity Analysis*", New York, John Wiley & Sons publishers.
- Sobol', I. M. (1990), "Sensitivity estimates for nonlinear mathematical models", *Matematicheskoe Modelirovanie*, 2: 112-118 (translated as: I. M. Sobol' (1993),

“Sensitivity analysis for non-linear mathematical models”, *Mathematical Modelling & Computational Experiment*, 1: 407-414.)

Young, P.C. (1984) *Recursive Estimation and Time-Series Analysis*, Springer.

Young, P.C. (1999) ‘Nonstationary time series analysis and forecasting’, *Progress in Environmental Science* 1 3-48.

Young, P.C., *Stochastic, Dynamic Modelling and Signal Processing: Time Variable and State Dependent Parameter Estimation*, Chapter from *Nonlinear and Nonstationary Signal Processing* edited by W. J. Fitzgerald et al., Cambridge University Press: Cambridge, 2000, 74-114.

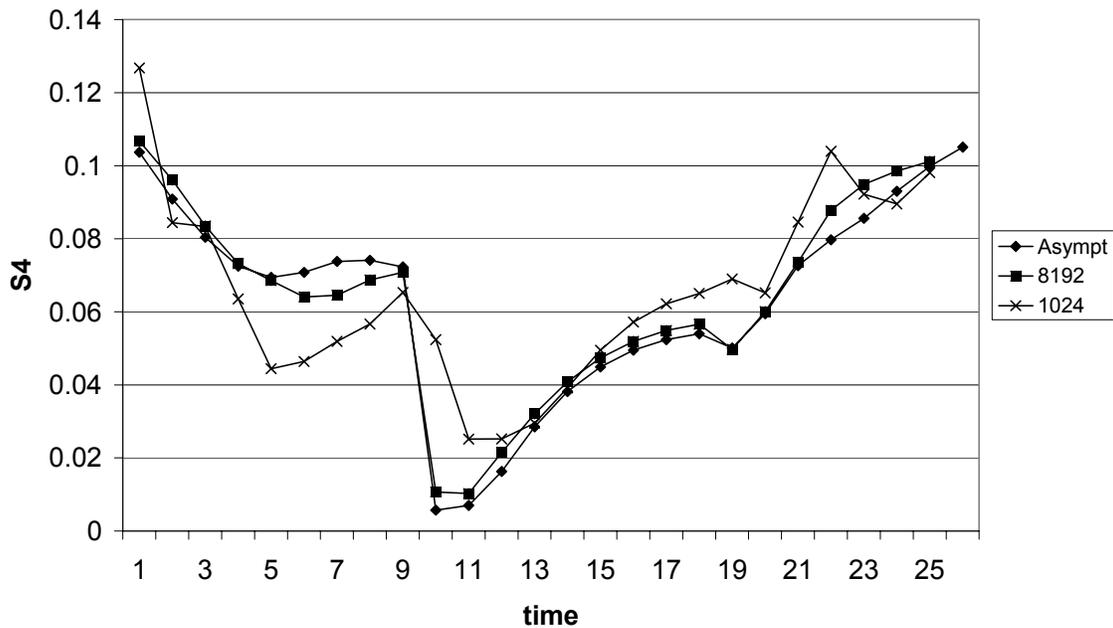


Figure 1. First order sensitivity index vs time for parameter  $v^{(1)}(X_4)$ .

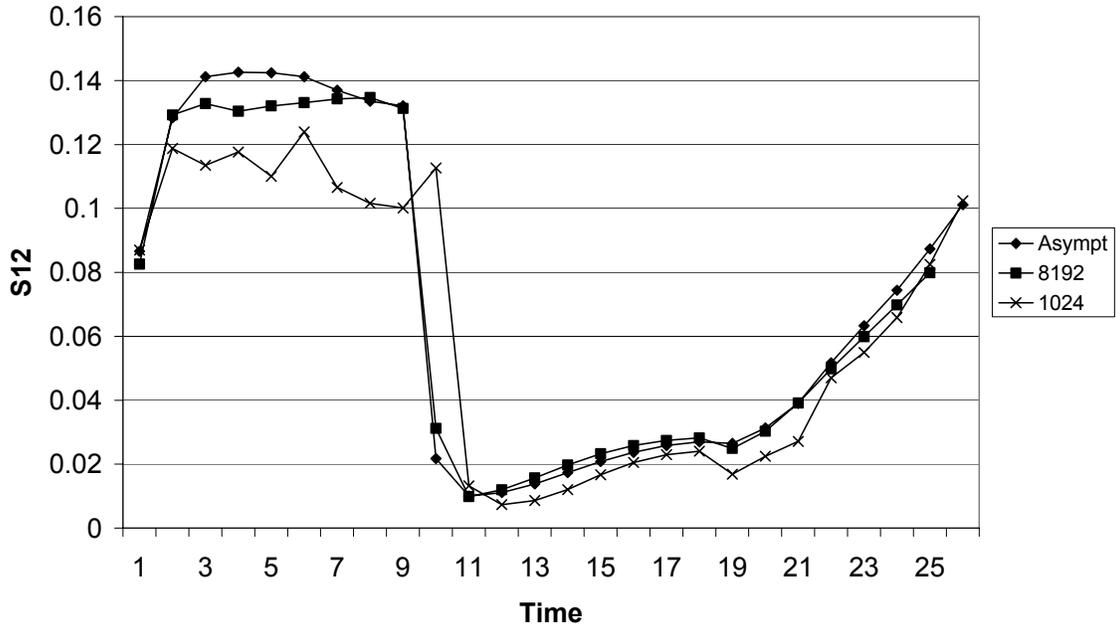


Figure 2. First order sensitivity index vs time for parameter W ( $X_{12}$ ).

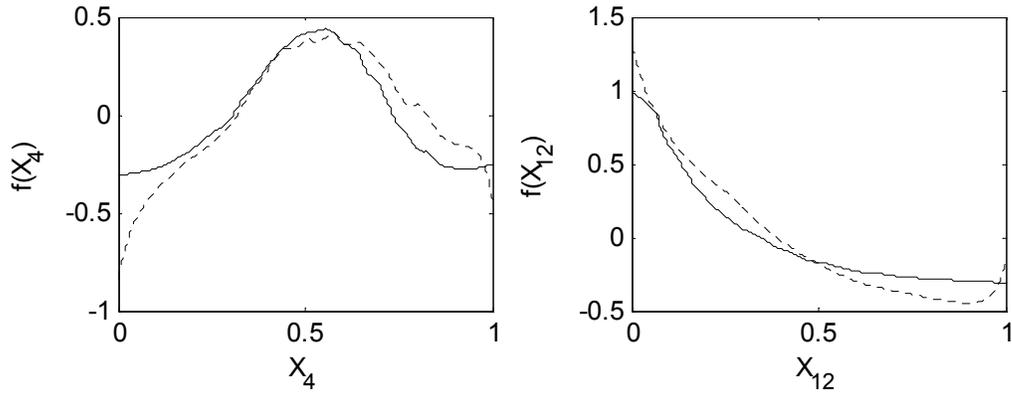


Figure 3. First order terms of the HDMR of the Level E for  $v^{(1)} (= X_4)$  and  $W (= X_{12})$ . Solid lines are for the total cost of 8192 runs; dotted lines for the total cost of 1024 runs (the scales of inputs and output are normalised).