Global Sensitivity Analysis of Stochastic Computer Models with Heteroscedastic GPs

Jack Kennedy

January 20, 2020

1 Intro

Here we aim to recreate the results of Marrel et al. (2012) using a Gaussian Process (GP) in place of the "true" model. Our GP model will be a Bayesian version of HetGP.

We aim to compute several sensitivity indices in this report. We will also explain some of the algorithms provided by Sobol (1993). Our focus here will be to estimate sensitivity indices of the Ishigami function; $f(x) = f(x_1, x_2, x_3) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4\sin(x_1)$. Where each $x_i \in [-\pi, \pi]$.

However, the Ishigami function is clearly deterministic, hence we modify it to produce a new function; $f(x, x_\varepsilon) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_\varepsilon^2\sin(x_1)$, with $x_\varepsilon \sim U(-\pi, \pi)$. Although the Ishigami function is simple to compute, in practice we will be working with much more complex systems which will be too expensive for a standard Monte Carlo sensitivity analysis. Hence, we will later replace f by an estimate $\hat{f}$ which will be a GP emulator.

Integrating out x_ε gives us the mean function for our stochastic Ishigami function;

$$Y_m(x) = \mathrm{E}(Y|x) = \left(1 + \frac{\pi^4}{50}\right)\sin(x_1) + 7\sin^2(x_2). \tag{1}$$

It is also straightforward to find the variance function;

$$Y_d(x) = \mathrm{Var}(Y|x) = \pi^8(900^{-1} - 2500^{-1})\sin^2(x_1). \tag{2}$$

Notice that Y_d only depends on x_1.

2 Sensitivity Indices & Their Computation

2.1 Sensitivity Indices

Sensitivity indices allow us to determine how important a group of variable are (or a single variable). They come in two forms, but both are variance based. For a model with p inputs

there are $2^p - 1$ such indices, and the sum of all of these indices is exactly one.

The first is

$$S_i = \frac{\mathrm{Var}(Y|X_i)}{\mathrm{Var}(Y)} \tag{3}$$

which is the proportion of variance that would be eliminated if we were to learn input(s) i.

The second type is

$$S_{T_i} = \sum_{J \supseteq i} S_J \tag{4}$$

which is the total sensitivity to group i. For example, in a three parameter problem, the total sensitivity to the first input is

$$S_{T_1} = S_1 + S_{12} + S_{13} + S_{123}. \tag{5}$$

A particularly important index is S_{T_ε}; this is the proportion of uncertainty induced by the random seed variable, including the interaction of the seed variable with the simulator's deterministic components.

2.2 Computation

Here we will outline the methods of Sobol (1993) to compute various sensitivity indices for a (stochastic) function f. In practice we might replace f by a cheap approximation, $\hat{f}$. We will also have an uncertainty distribution over the inputs. I.e. $x \sim G(x)$, where G is the uncertainty distribution. Further, suppose we can split x into groups $x = (x^1, x^2)$. We can then think about the proportion of variation explained by each group of variables, but a "group" might be a single variable (e.g. x_1 or x_ε - uncontrollable variable.

We first want to compute the mean value of the mean function, $f_0 = E(Y_m)$

$$\begin{aligned} f_0 &= \int Y_m(x) \mathrm{d}G(x) \\ &\approx \frac{1}{N} \sum_{j=1}^{N} \hat{Y}_m(x_j) \end{aligned}$$

where N is a large number, x_j are iid samples from $G(x)$ and $\hat{Y}_m$ is an estimate of the mean function of f. When we write x_j, we will usually mean a sample from $G(x)$.

The variance of Y_m (i.e. the uncertainty in Y_m induced by $G(x)$). This is simply

$$V = \mathrm{Var}(Y_m) \tag{6}$$
$$= \int Y_m(x)^2 \mathrm{d}x - f_0^2 \tag{7}$$
$$\approx \frac{1}{N}\sum_{j=1}^{N} \hat{Y}_m(x_j)^2 - f_0^2. \tag{8}$$

We can find $V_1 = Var(Y|X_1)$ in a similar way:

$$V = \mathrm{Var}(Y_m|X_1) \tag{9}$$
$$= \int Y_m(x_1)^2 \mathrm{d}x_1 - f_0^2 \tag{10}$$
$$\approx \frac{1}{N}\sum_{j=1}^{N} \hat{Y}_m(x_1, x_2)\hat{Y}_m(x_1, x_j^{2\prime}) - f_0^2. \tag{11}$$

Here, y_j is a draw from $G_1(x^1)$, the marginal distribution of x^1, and then $x_j^2, x_j^{2\prime}$ are a pair of independent draws from $G_{2|1}(x^2|x^1 = x_j^2)$

3 Application to The Ishigami Function

For the Ishigami function we treat x_3 as the random input. Based on a training sample of 500 data points (no replication) we build a HetGP emulator. We obtain the following estimates for sensitivity measures:

	S_1	S_2	S_{12}	S_{T_3}
Exact Values	0.314	0.442	0	0.244
Estimates	0.318	0.461	0.004	0.228
|Percentage Error|	1.27	4.30	–	6.56

4 Conclusions

We see that $S_{T_3} = S_{T_\varepsilon}$ is estimated to be 0.23, hence, 23% of output uncertainty is due to the stochastic nature of the simulator. We see $\hat{S}_2 = 0.461$ hence a large proportion of output uncertainty is attributed to x_2, similarly, a large (but smaller) proportion of the uncertainty is attributed to the variation in x_1. We see that S_{12} is small so it is likely that there is a very weak, or possibly a non-existent, interaction between the two controllable inputs. We also see that the proportion of variance explained by the controllable variables is estimated by $1 - \hat{S}_{T_\varepsilon} = 0.772$, hence the controllable variables dominate the output uncertainty.

5 Application to Athena model

5.1 Sensitivity Indices

We now apply the above methodology to the Athena model (Zitrou et al. 2013, 2016). We will use pre-exisiting emulators of the Athena model to perform the computation. We will try a HetGP emulator but also a Stochastic Multilevel (SML) emulator for sensitivity index estimation. Note that the Athena model produces values in the range $[0, 1]$, so we in emulate the probit Availability. Hence, sensitivity indices will be for the probit of the output, rather than the output itself.

For the distribution over the unknown parameters, we assumed $x_i^* \sim \mathcal{U}(-1.7, 1.7)$ where each x_i^* is the studentised version of the original inputs. Future work would involve producing a distribution over x which is based on Expert opinion.

Here we estimate S_i for $i = 1, \ldots, 6$ and also S_{T_ε}. The estimates are based on a Monte-Carlo procedure with a sample size of $N = 10^5$.

	S_1	S_2	S_3	S_4	S_5	S_6	S_{T_3}
HetGP	0.013	0.099	0.039	0.108	0.254	0.151	0.266
SML	0.005	0.080	0.063	0.131	0.285	0.213	0.136

From Section 5.1 we see that, the (estimated) S_i have the same ordering ($S_5 > S_6 > S_4 > S_2 > S_3 > S_1$). We also see that $S_{T_\varepsilon}^{\text{HetGP}} > S_{T_\varepsilon}^{\text{SML}}$, indicating that the total uncertainty due to the random seed variable is estimated to be larger via HetGP. This could however be larger due to an ill–fitting mean function, and hence deviations from the mean function are being intepreted as noise.

The Sobol sensitivity should sum exactly to 1. In our case we see that $\sum S_i + S_{T_\varepsilon}$ is 0.930 for HetGP 0.914 for the SML emulator, suggesting that interactions between the controllable variables are present. The contribution to input uncertainty amongst interaction terms is quite small, but estimated to be slightly larger by SML emulation that HetG. The total contribution to output uncertainty induced by interactions amongst controllable variables is estimated to be less that 10% in both cases.

5.2 Interaction Effects

We should also investigate which are the largest interactions, should they exist. We, for now, will only do this for the SML emulator.

To find the interaction indices, we must first find $V_{ij} = \text{Var}\left\{\text{E}(Y|x_i, x_j)\right\}$. This is the uncertainty induced by the linear and interaction effect between x_i and x_j. We then have

$$S_{ij} = \frac{V_{ij} - V_i - V_j}{\text{Var}(Y)}. \tag{12}$$

Using a simple Monte-Carlo estimation technique, with $N = 10^5$ samples, we see that the following are the largest interaction terms:

$$100\hat{S}_{23} = 1.96$$
$$100\hat{S}_{34} = 1.24$$
$$100\hat{S}_{36} = 0.80$$

As expected, these interaction effects are quite small. It is interesting to see that all of them have the 3 in their index - this variable is the cable repair time.

5.3 Main effect Inference

Here we aim to estimate

$$\mathrm{E}(Y|x_i) = \int_{\mathcal{X}_{-i}} \eta(\mathbf{x}) \mathrm{d}G_{-i|i}(X_{-i}|X_i). \tag{13}$$

Here, $\mathrm{E}(Y|x_i)$ is the expected response from the simulator if we were to learn the input $X_i = x_i$ (or indeed, this can be generalised if X_i is a group of variables). We compute this for a grid of values over each i. In practice, we replace $\eta(\cdot)$ with an emulator (we use the same as constructed earlier). Since the Athena model is stochastic, we replace $\eta(\cdot)$ with an estimate of the simulator mean and use Monte Carlo methods to estimate the integrals. More explicitly,

$$\mathrm{E}^*\{E(Y|x_i)\} \approx \frac{1}{N} \sum_{j=1}^{N} Y_m(x_{(j)}) \tag{14}$$

where $x_{(j)}$ are iid draws from $G_{-i|i}(x_i|x_{-i})$

However, we will also want estimates of uncertainty about the expected mean functions. There are two sources of uncertainty in emulators of stochastic functions; the epistemic (extrinsic) uncertainty in the mean and variance due to lack of knowledge of the simulator but also the aleatory (intrinsic) uncertainty due to the stochastic nature of the simulator. Taking the epistemic uncertainty into account will give us an interval for the mean value given x_i. Taking both epistemic and aleatory uncertainty into account gives us an interval which represents likely values for a new run of the simulator given x_i, as seen below.

In Figure 1 we see that the effect curves are similar for both HetGP and SML. However, the uncertainty bands are quite different. In this case, the uncertainty bands are calculated in a similar way to the main effects, we integrate the variance over $G_{-i|i}(X_{-i}|X_i)$. Futher, these uncertainty bands represent out "total" uncertainty in prediction - they are a combination of the uncertainty in the mean function but also the stochastic nature of the simulator, so give us (approximate) 95% probability bands for a new simulator run, if we were to learn the value of x_i. We see that variables 4 — 6 are the most influential, which suggests that the best way to have a windfarm with high availability is to have well functioning components.

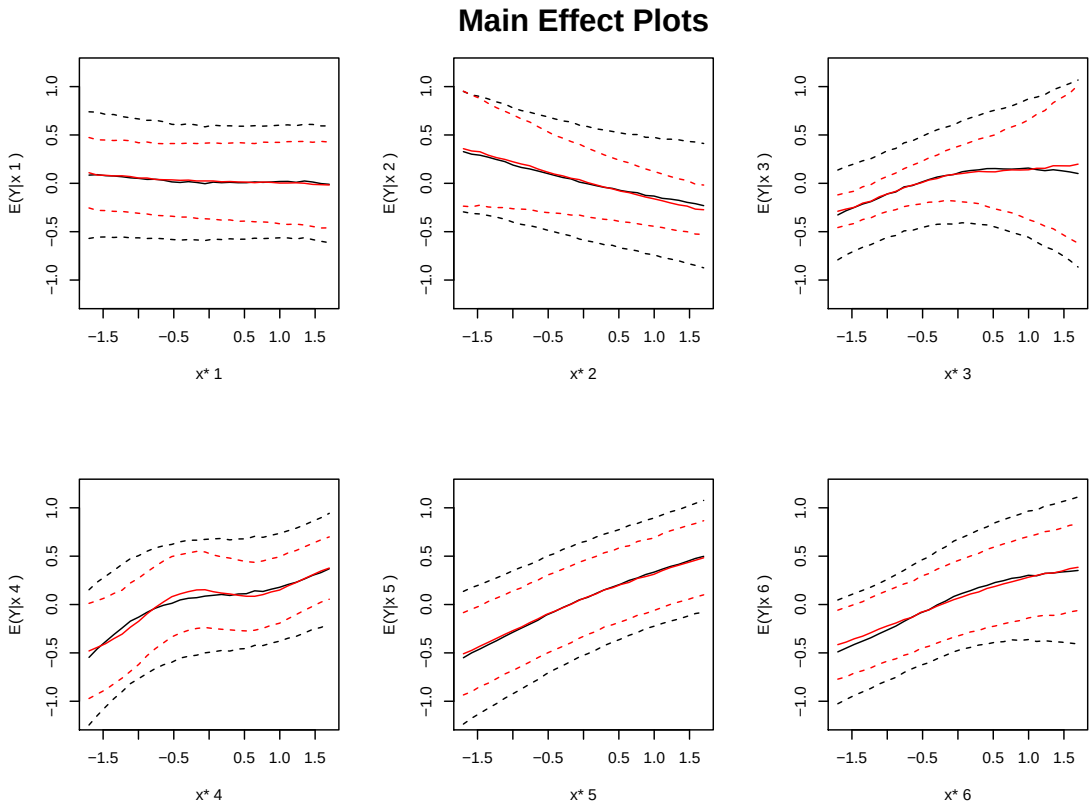


Figure 1: Main effect curves (solid) ±2 standard deviations (dashed). Black represents HetGP and red is SML emulation.

6 Choice of Prior

To perform the sensitivity analysis, we must specify a distribution, G. As mentioned by **?**, the model inputs must be deemed as independent in this screening process to ensure a unique ANOVA decomposition of the simulator. We will now perform an investigation into the choice of G subject to a few constraints:

- Each choice of G has the same mean.
- Each choice of G is symmetric.
- Each choice of G has a similar 0.05% and 99.5% quantiles.

The analysis above used uniform priors over the range of the observed data. However, using a distribution with a hard cut-off may not be satisfactory. For instance, the uncertain input in question could plausibly take any value on the real line, but we just believe that it is likely to be between two limits. We will experiment with a choice of two symmetric, unimodal choice of G. First we will try $G(x_i^*) \sim \mathcal{N}(0, 0.65^2)$. We will also try $G(X_i^*) \sim 0.536t_{10}$. These were chosen since they give $P(-1.7 < x_i^* < 1.7) = 0.99$. The uniform priors from before have the property $P(-1.7 < x_i^* < 1.7) = 1$, so these distributions have similar quantiles. They are also all symmetric, unimodal and have mean 0. Therefore, each of these priors have many similar summaries. They could all be deemed as an adequate (coarse) representation of beliefs by an expert.

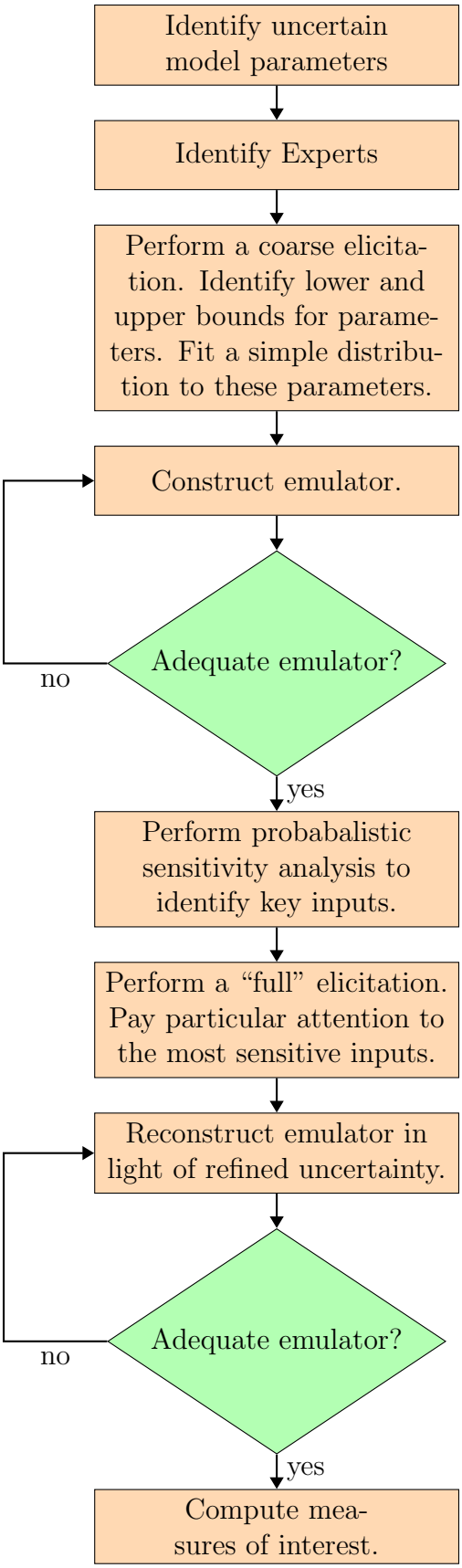
We then compute the first order sensitivity idices and S_{T_ε} via a simple MC method. We can then compare the values of the indices to gauge the relative importance of each variable under different prior specifications.

G	$100\hat{S}_1$	$100\hat{S}_2$	$100\hat{S}_3$	$100\hat{S}_4$	$100\hat{S}_5$	$100\hat{S}_6$	$100\hat{S}_{T_\varepsilon}$
$\mathcal{U}$	0.72	8.18	5.96	13.88	29.17	22.20	13.81
$\mathcal{N}$	0.36	7.87	8.56	7.24	25.96	27.12	15.85
t_{10}	0.50	7.82	9.12	6.28	25.25	28.41	16.96

Table 1: Estimated first order sensitivity indices and S_{T_ε} for three simple choices of G.

From Table 1 we see that for the t_10 and Normal priors, $S_6 > S_5 > S_3 > S_2 > S_4 > S_1$ whereas for the uniform prior $S_5 > S_6 > S_4 > S_2 > S_3 > S_1$. The only real agreement between these two formulations is that they both rank S_1 as the least informative first order effect. They all rank S_5 and S_6 as the two most important first order effects. The row sums for the above table are 93.92, 92.96 and 94.34 respectively (top - bottom). Therefore, in each case the interactions between controllable variables have a similar effect; of the order of $5\% - 7\%$.

7 Flowchart


Identify uncertain model parameters
Identify Experts
Perform a coarse elicitation. Identify lower and upper bounds for parameters. Fit a simple distribution to these parameters.
Construct emulator.
Adequate emulator?
no
yes
Perform probabalistic sensitivity analysis to identify key inputs.
Perform a "full" elicitation. Pay particular attention to the most sensitive inputs.
Reconstruct emulator in light of refined uncertainty.
Adequate emulator?
no
yes
Compute measures of interest.

References

Marrel, A., Iooss, B., Da Veiga, S. & Ribatet, M. (2012), 'Global sensitivity analysis of stochastic computer models with joint metamodels', *Statistics and Computing* **22**(3), 833–847.

Sobol, I. M. (1993), 'Sensitivity estimates for nonlinear mathematical models', *Mathematical modelling and computational experiments* **1**(4), 407–414.

Zitrou, A., Bedford, T. & Walls, L. (2016), 'A model for availability growth with application to new generation offshore wind farms', *Reliability Engineering and System Safety* **152**(C), 83–94.

Zitrou, A., Bedford, T., Walls, L., Wilson, K. & Bell, K. (2013), Availability growth and state-of-knowledge uncertainty simulation for offshore wind farms, *in* '22nd ESREL conference 2013'.
URL: *https://strathprints.strath.ac.uk/45377/*