Sensitivity Analysis in Epidemiological Modelling

Jeremy Oakley

Department of Probability & Statistics
University of Sheffield

Acknowledgements

The first case study in the MUCM project. Case study team:

- John Paul Gosling (Food and Environment Research Agency)
- Hugo Maruri-Aguilar (LSE)
- Alexis Boukouvalas (Aston)

with contributions from Leo Bastos (Sheffield), Dan Conford (Aston), Thierry van Effelterre (GSK) and Henry Wynn (LSE).

- First version of the rotavirus model was shared by GlaxoSmithKline for MUCM Case study

Introduction

- Simulator models spread of rotavirus in a population, and effect of vaccination on incidence
- Various uncertain simulator inputs, which induce uncertainty in model outputs
- Can we (efficiently) identify the most 'important' uncertain simulator inputs?

Identifying important (and unimportant) inputs useful for

1. Understanding how the model works
2. Prioritising model development
3. Prioritising data collection

Rotavirus background

- Leading cause of gastroenteritis, major cause of diarrhoea-related hospitalizations and deaths among young children worldwide
- Approx 611,000 children under five die from rotavirus infection each year in developing countries
- Average of 37 deaths per year in the United States between 1993 and 2003.
- Virus transmitted from person-to-person, mainly by faecal-oral route
- First infection often (but not always) symptomatic
- First and subsequent rotavirus infections progressively build up natural immunity

The simulator

- Simulator models incidence of rotavirus in a population before and after a vaccine is administered to a proportion of the infant population
- Main objective of model: project population-level impact of vaccination, including direct and indirect “herd protection” effects
- Deterministic compartmental model
- 20 simulator inputs, e.g. transmission rates between age groups, reduction in risk following each infection
- Simulator outputs: time series of rotavirus incidence for six age groups following vaccination programme
  - Four time points of interest, so 24 outputs in total
  - We report on analysis of individual outputs (e.g. no. of infections in age group 0-1 years, 1 year after vaccination programme)

Sensitivity Analysis

- Represent our simulator by the function $y = \eta(\mathbf{x})$, with $\mathbf{x} = (x_1, \ldots, x_d)$
- What do we mean by an ‘influential input’?
- How does y vary as elements of $\mathbf{x}$ vary?

One-way sensitivity analysis

- Write $\mathbf{x} = (x_i, \mathbf{x}_{-i})$ with $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.
  1. Fix $\mathbf{x}_{-i}$ at some ‘central’ value $\mathbf{x}^*_{-i}$
  2. Choose lower and upper limits $x_i^{(l)}$ and $x_i^{(u)}$
  3. Evaluate $\eta(x_i^{(l)}, \mathbf{x}^*_{-i})$ and $\eta(x_i^{(u)}, \mathbf{x}^*_{-i})$

- Difficulties with the one-way approach:
  1. What if η is nonlinear in x_i?
  2. What should the limits $x_i^{(l)}$ and $x_i^{(u)}$ be?
  3. Does the choice of $\mathbf{x}^*_{-i}$ matter?
  4. Are $(x_i^{(l)}, \mathbf{x}^*_{-i})$ and $(x_i^{(u)}, \mathbf{x}^*_{-i})$ plausible input values?

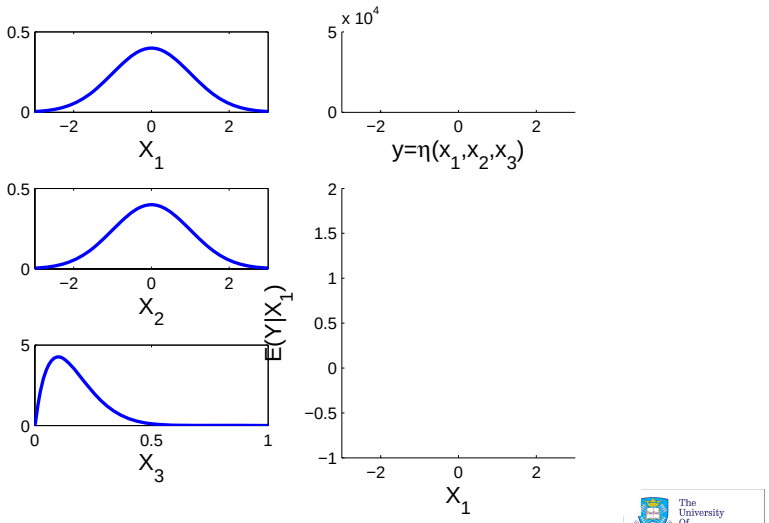
Variance-based sensitivity analysis

- Suppose we are interested in $Y = \eta(\mathbf{X})$ for some uncertain ‘true’ input $\mathbf{X}$?
- $p(\mathbf{X})$ represents subjective uncertainty about true input $\mathbf{X}$
- Can investigate how elements of $\mathbf{X}$ contribute to uncertainty in Y: those with largest contribution classed as most ‘influential’

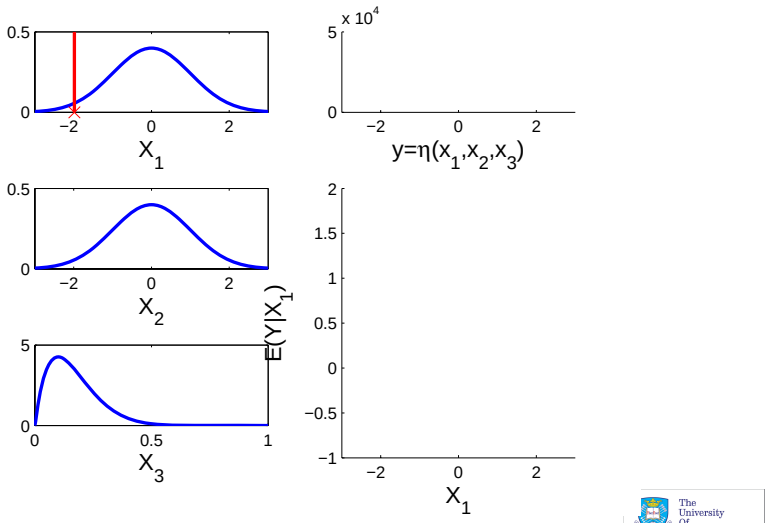
Example

- Three input function $y = \eta(x_1, x_2, x_3)$
- We consider uncertainty in $Y = \eta(X_1, X_2, X_3)$
- (X_1, X_2) bivariate normal distribution with correlation 0.5
- X_3 independent Beta distribution

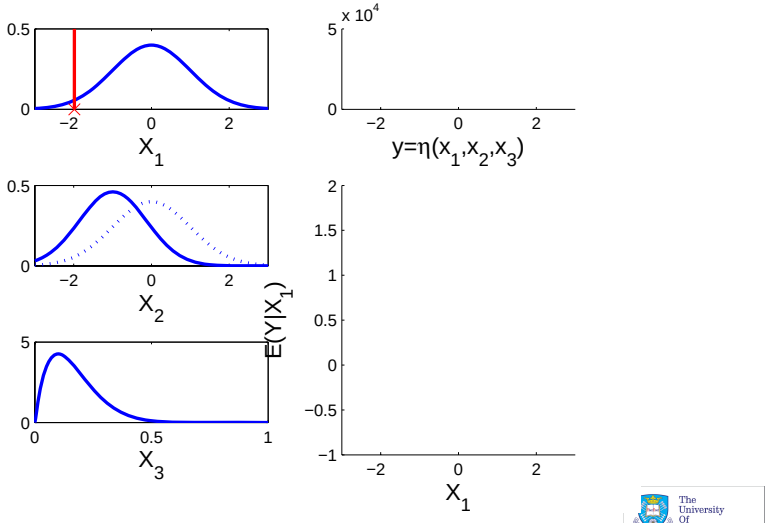
Distributions of the three inputs



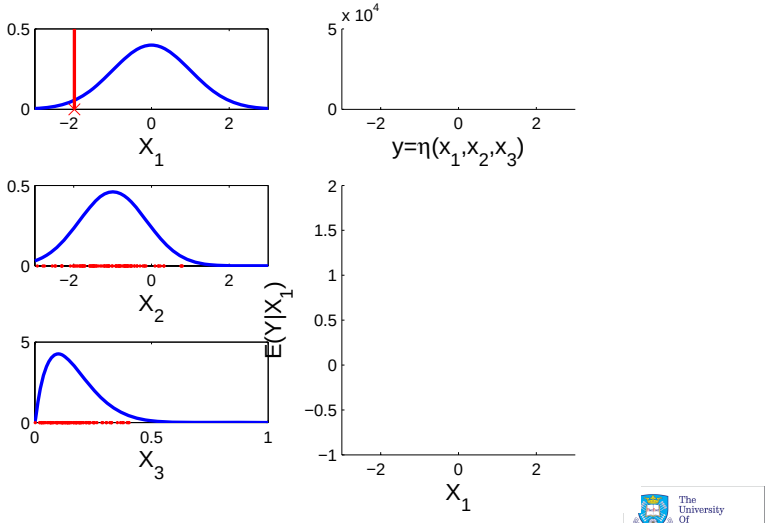
Choose value of X_1



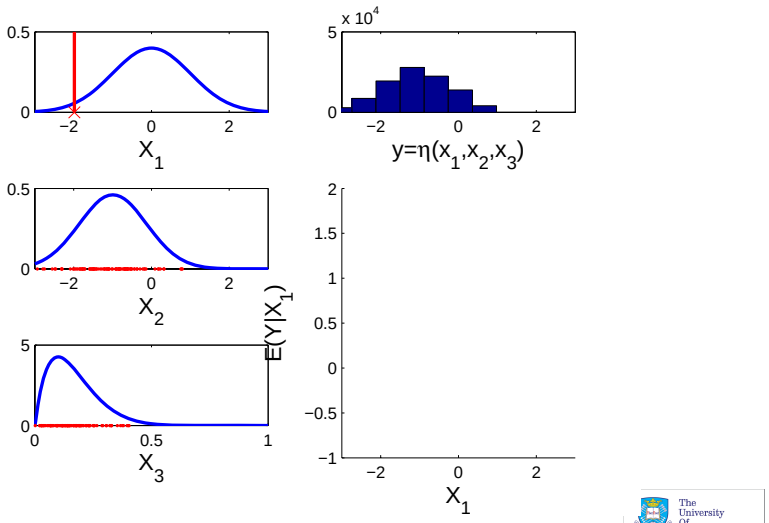
Update distributions of other inputs given X_1



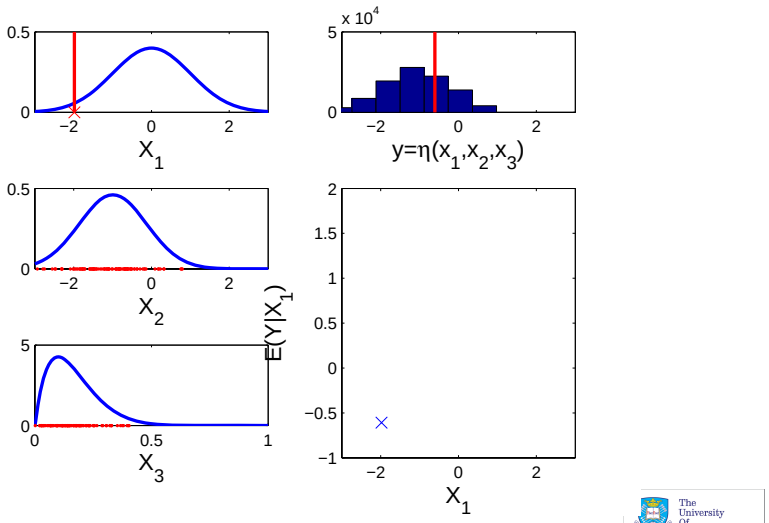
Sample from $p(X_2, X_3|X_1)$



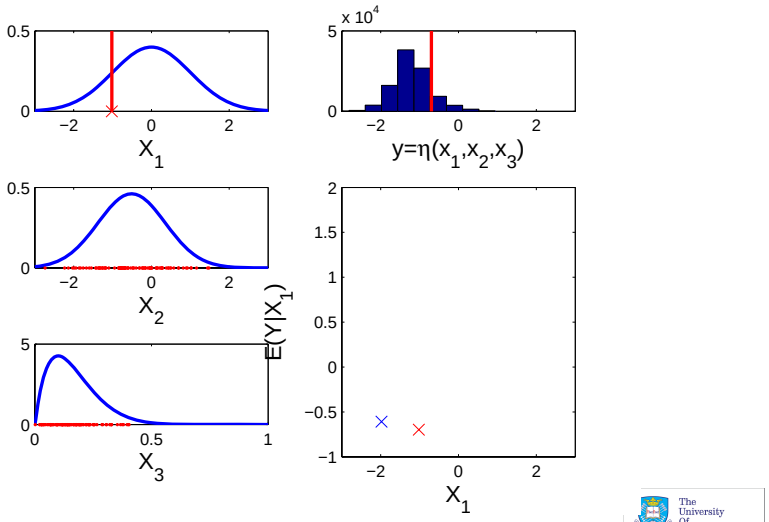
Run model at each set of sampled inputs



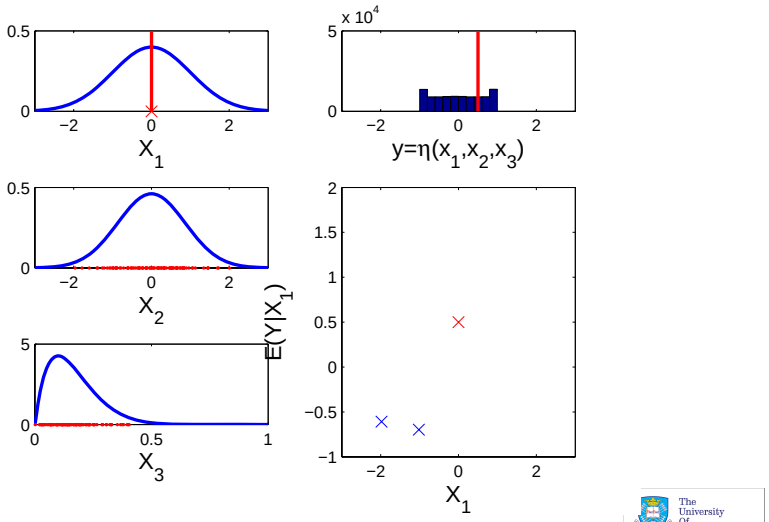
Obtain the mean output $E(Y|X_1)$



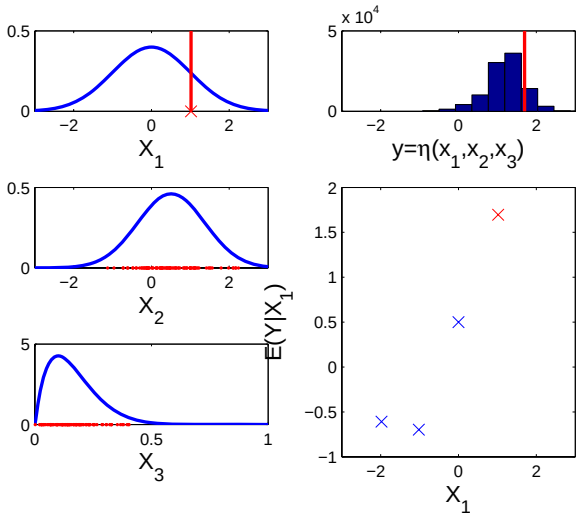
Repeat for new value of X_1



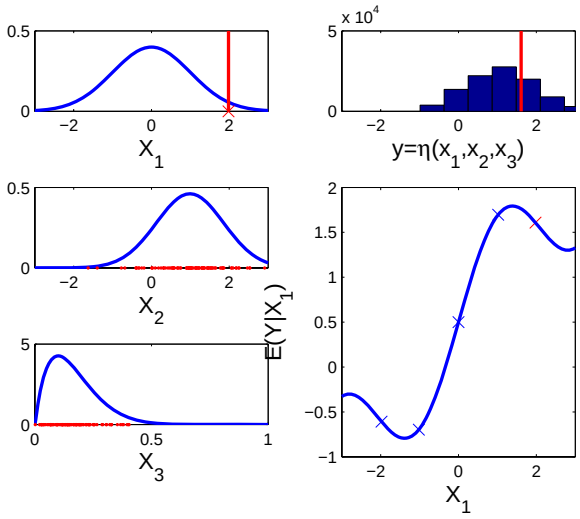
Repeat for new value of X_1



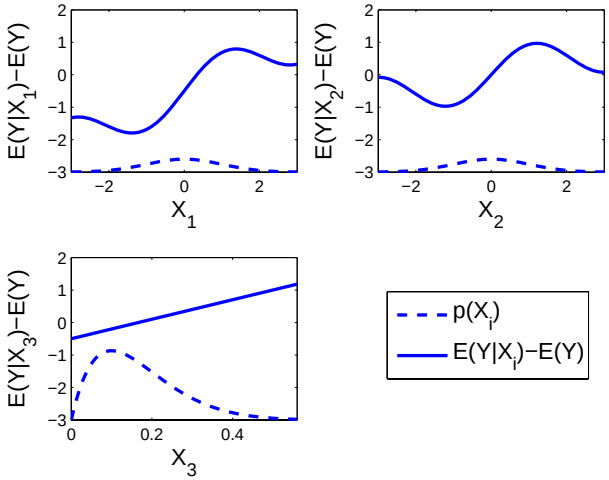
Repeat for new value of X_1



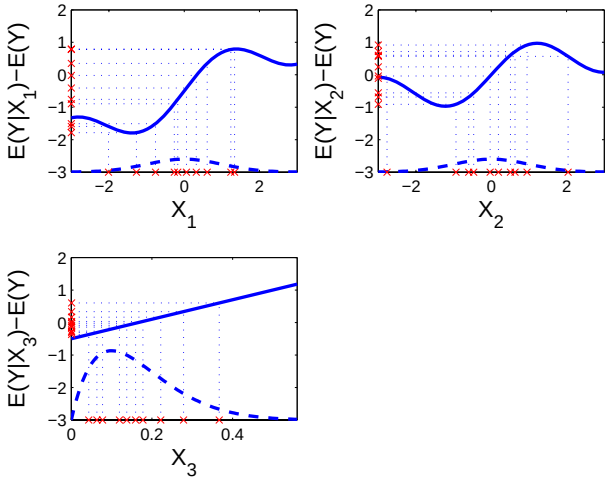
Obtain $E(Y|X_1)$ as function of X_1



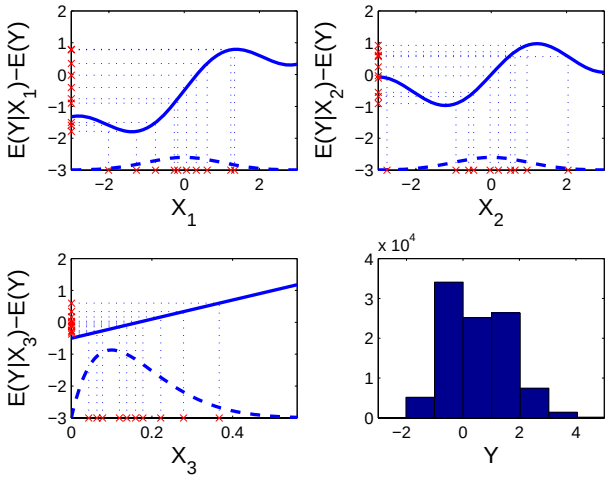
$E(Y|X_i)$ and $p(X_i)$



Sample from $p(X_i)$, observe variability in $E(Y|X_i)$



Compare with variability in Y



Variance based sensitivity analysis

- Main effect plot: $E(Y|X_i) - E(Y)$ against X_i

Main effect variance V_i and index S_i

$V_i = Var_{X_i}\{E(Y|X_i)\}$, with main effect index $S_i = V_i/Var(Y)$

- Note expected reduction in variance obtained by learning X_i:

$$Var(Y) - E_{X_i}\{Var(Y|X_i)\} = Var_{X_i}\{E(Y|X_i)\}.$$

Total effect variances and indices: an example

Consider $Y = \eta(\mathbf{X}) = X_1 + X_2 + X_1X_3$, with $X_1, X_2, X_3 \sim N(0,1)$
$Var(Y) = 3$

$$Var_{X_1}\{E(Y|X_1)\} = 1 \qquad Var_{X_2}\{E(Y|X_2)\} = 1 \qquad Var_{X_3}\{E(Y|X_3)\} = 0$$

Additional variance due to interaction between X_1, X_3:

$$Var_{X_1,X_3}\{E(Y|X_1,X_3)\} - Var_{X_1}\{E(Y|X_1)\} - Var_{X_3}\{E(Y|X_3)\} = 1$$

Total effect variance V_{T_i} and index S_{T_i}

Variance of main effect of X_i + additional variance due to all interactions involving X_i, with total effect index $S_{T_i} = V_{T_i}/Var(Y)$

Total effect variances and indices

$Y = \eta(\mathbf{X}) = X_1 + X_2 + X_1X_3$, with $X_1, X_2, X_3 \sim N(0,1)$
$Var(Y) = 3$

$$V_{T_1} = 2 \qquad V_{T_2} = 1 \qquad V_{T_3} = 1$$

1. $V_{T_i} = E\{Var(Y|\mathbf{X}_{-i})\}$
2. A total effect index close to 0 identifies an 'unimportant' input
3. Independence between inputs required

Computation

- Main/total effect indices can be computed using Monte Carlo...
- ...but large number of model runs required
- Simple Monte Carlo not practical for computationally expensive models
- Emulators can be used for complex models
- For certain emulator modelling choices, can quickly estimate $E(Y = \eta(X)|X_i)$ as well as $y = \eta(\mathbf{x})$

Input screening

- GP emulators difficult to implement for simulators with large numbers of inputs
  - Variance-based SA using the emulator will identify unimportant inputs...
  - ...but need to identify unimportant inputs before building the emulator
- We consider cruder SA methods to first screen out inactive inputs

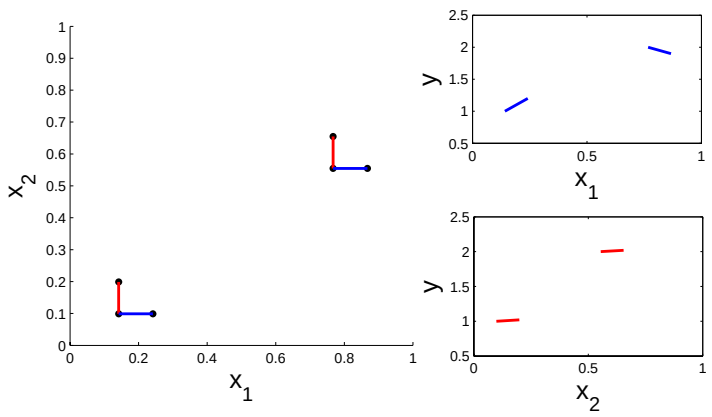
Input screening

- Variation on Morris (1995) one-at-a-time screening method
- Involves calculating 'elementary effects' (partial derivatives) for each input at various points, e.g.,

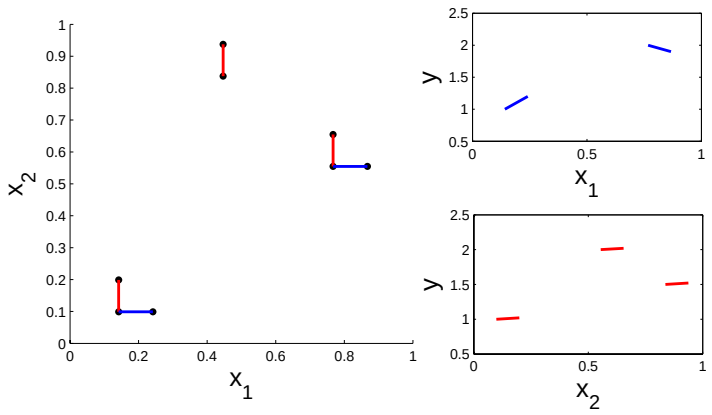
$$\frac{\eta(x_{1,i} + h, x_2, \ldots, x_d) - \eta(x_{1,i}, x_2, \ldots, x_d)}{h},$$

  for $i = 1, 2, \ldots$
- Large variance (or mean) of elementary effect identifies an active input
- Space-filling blocks chosen both to estimate elementary effects and support emulator construction

Simulator design for input screening

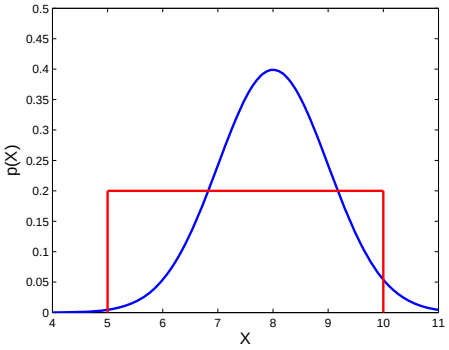


Simulator design for input screening



Input screening

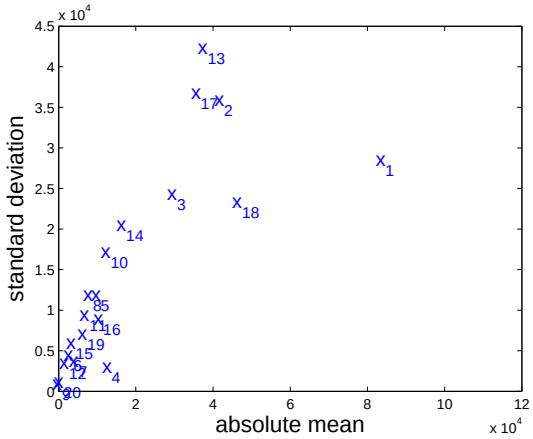
Screening exercise conducted twice



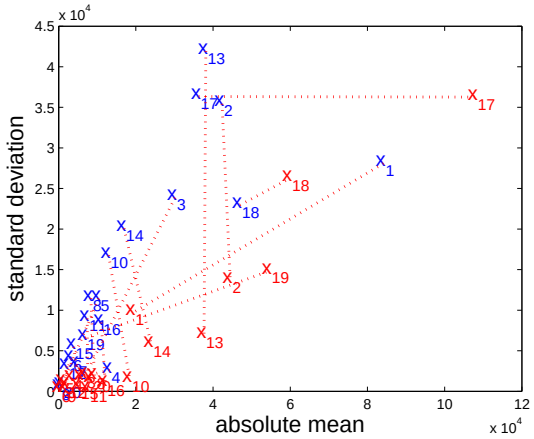
1. with uniform ranges chosen for each uncertain input
2. with formally elicited input distributions

42 simulator runs used in each case, 24 outputs considered individually

Screening results with uniform input distributions



Screening results with formally elicited distributions

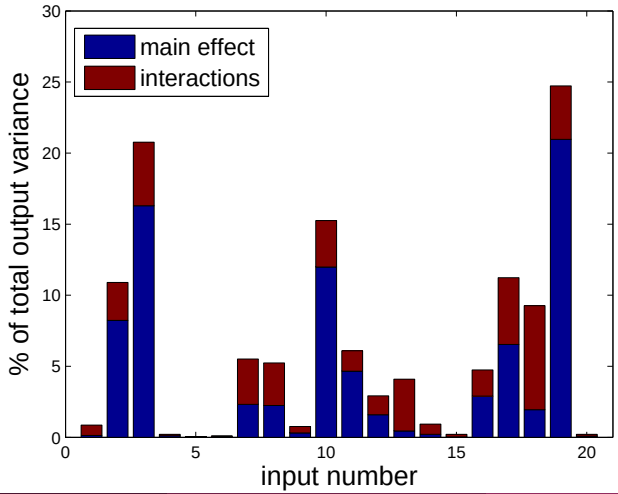


Emulators

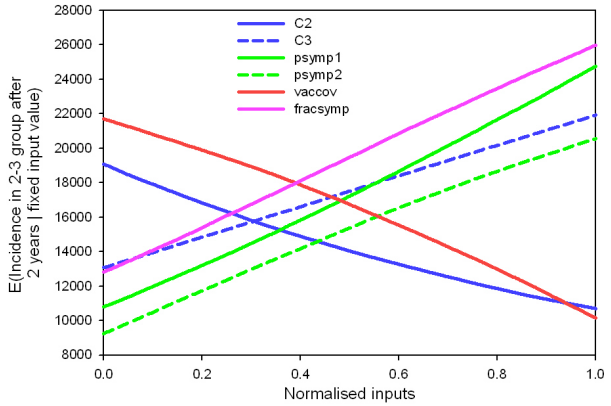
- All inputs found to be ‘active’ for at least one output
  - Emulator built using all 20 inputs
- Emulator built with 200 simulator runs
- Main and total effect indices calculated for each input

Variance based sensitivity analysis

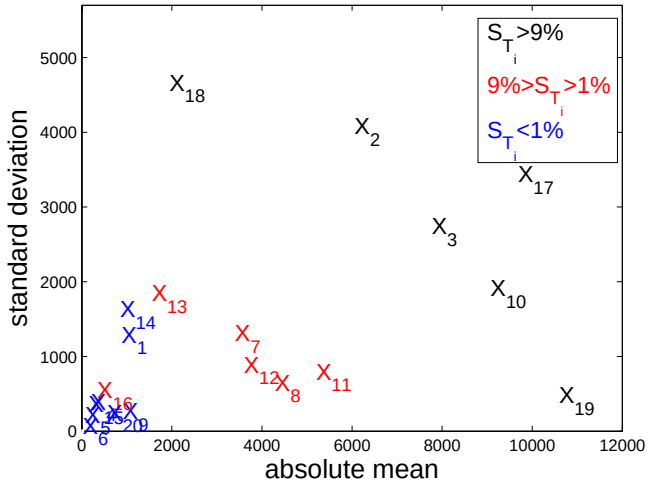
Analysis for an individual output: no. of infections in 2-3 age group after 2 years



Main effects plots



Correspondence between screening results and variance based sensitivity analysis



Summary

- Important inputs identified by their contribution to the output variance. Simple screening approaches also useful here
- Emulators used for efficient computation
  - 342 runs with 20 uncertain inputs
  - Previous GSK analysis 8200 runs with 9 uncertain inputs
- Important to represent input uncertainty carefully