SAND2021-13811C

Sandia National Laboratories

Exceptional service in the national interest

Empirical Sensitivity Analysis: Extensions and Open Questions

Justin Winokur, PhD
jgwinok@sandia.gov

Sandia National Laboratories

Verification & Validation, Uncertainty Quantification, and Credibility Processes Department

Albuquerque, NM

UC Davis/Sandia 2021 Sandia Fall Symposium
November 3, 2021

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Background & Motivation

Sensitivity Analysis is the process of identifying dominant parameters; usually with the goal of excluding some from future analysis or to prioritize additional analysis activities.

- Precision is *not* of paramount importance!

Global Sensitivity Analysis (GSA) is powerful and robust but can be very expensive!

- Surrogate-Driven GSA requires more complete knowledge of the model and relies on a potentially inaccurate surrogate.
  - More complex
  - Hard to trace sources of error including surrogate model form error
  - Can be slow or difficult to fit

Empirical Sensitivity Analysis:

- Based solely on data with *minimal* assumptions
- Based on *observation*, not *interrogation*
  - Can use existing data without ability to add more.
- No need for prior knowledge of probability spaces
  - Assume the data *describes* the input probability
- Intended to be fast and robust, even at the cost of accuracy
- Use Sobol Sensitivity Indices: i.e. percent of the variance is due to each parameter

Sobol' Sensitivity Analysis

Sobol' Decomposition: Decompose $\mathcal{L}^2$ function $f(x_1, \dots, x_n)$ into 2^n functionals orthogonal with respect to covariance such that $Cov[f_i f_j] = \mathbb{V}[f_i]\delta_{ij}$

$$f(x_1, \dots, x_n) = f_\emptyset + \sum_{1 \le i \le n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_{ij}) + \cdots = \sum_{\boldsymbol{d} \in \mathcal{P}(\{x_1, \dots, x_n\})} f_{\boldsymbol{d}}(\boldsymbol{x_d})$$

$\mathcal{P}$: Powerset. Set of all subsets including itself and the empty set

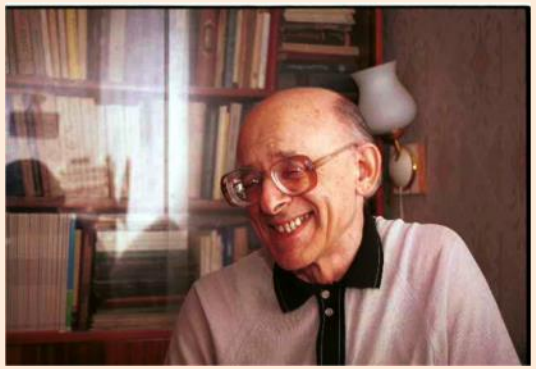
Where $f_{\boldsymbol{d}}(\boldsymbol{x_d}) = \mathbb{E}[f|\boldsymbol{x_d}] - \sum_{\boldsymbol{k} \in \mathcal{P}(\boldsymbol{d}) \setminus \boldsymbol{d}} f_{\boldsymbol{k}}(\boldsymbol{x_k})$

Variance of Sobol' Decomposition: Due to orthogonality, variance operator can move into the summation:

$$\mathbb{V}[f(x_1, \dots, x_n)] = \mathbb{V}\left[\sum_{\boldsymbol{d} \in \mathcal{P}(\{x_1, \dots, x_n\})} f_{\boldsymbol{d}}(\boldsymbol{x_d})\right] = \sum_{\boldsymbol{d} \in \mathcal{P}(\{x_1, \dots, x_n\})} \mathbb{V}[f_{\boldsymbol{d}}(\boldsymbol{x_d})]$$

Example: Simplify Notation: $\mathbb{V}[f_{\boldsymbol{d}}(\boldsymbol{x_d})] = \mathbb{V}_{\boldsymbol{d}}$ and note $\mathbb{V}_\emptyset = \mathbb{V}[constant] = 0$

$$\mathbb{V}[f(x, y, z)] = \mathbb{V}_x + \mathbb{V}_y + \mathbb{V}_z + \mathbb{V}_{xy} + \mathbb{V}_{xz} + \mathbb{V}_{yz} + \mathbb{V}_{xyz}$$



Ilya M. Sobol' (with an apostrophe)

Sobol' Indices

Recall there are 2^n function in a Sobol' Decomposition.

Simplify into Main (S) and Total (T) contributions. Main is sometimes known as "First-Order" or "Primary"

Recall

$$\mathbb{V} = \mathbb{V}_x + \mathbb{V}_y + \mathbb{V}_z + \mathbb{V}_{xy} + \mathbb{V}_{xz} + \mathbb{V}_{yz} + \mathbb{V}_{xyz}$$

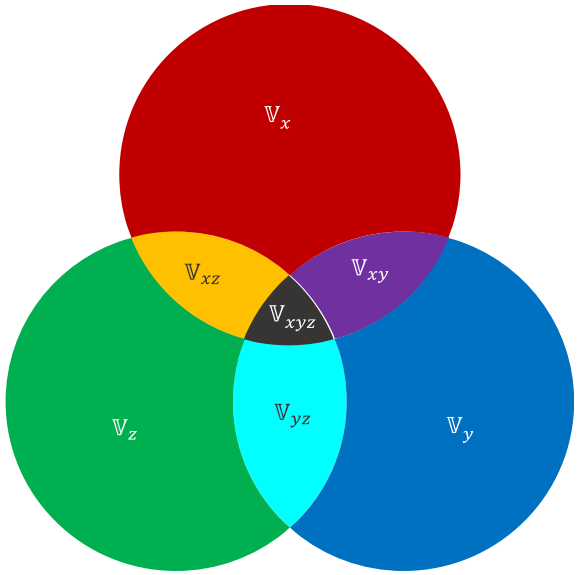
Define:

$$S_{\boldsymbol{d}} = \frac{\text{Variance due } \textbf{exclusivly} \text{ to dimension(s) in } \boldsymbol{d}}{\text{total variance}} = \boxed{\frac{\mathbb{V}\left[\mathbb{E}[f|\boldsymbol{d}]\right]}{\mathbb{V}}}$$

$$T_{\boldsymbol{d}} = \frac{\text{Variance due to dimension(s) in } \boldsymbol{d} \text{ and } \textbf{any} \text{ interactions}}{\text{total variance}}$$

Takeaway:
The "Main Sobol' Index" is the fraction of the variance due *only* to a parameter. The "Total Sobol' Index" includes interactions.

The numerator of the Main Index is the *variance of the conditional expectation*



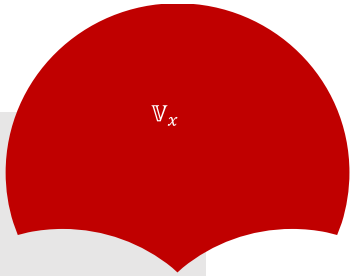
Example:

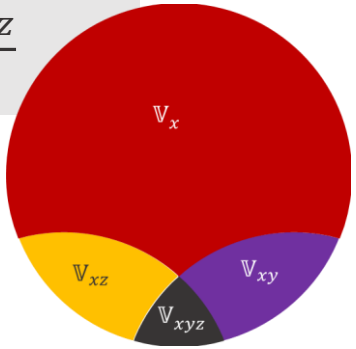
$$S_x = \frac{\mathbb{V}_x}{\mathbb{V}} = \frac{\mathbb{V}\left[\mathbb{E}[f|x]\right]}{\mathbb{V}}$$

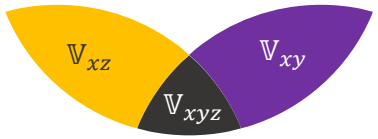
$$T_x = \frac{\mathbb{V}_x + \mathbb{V}_{xy} + \mathbb{V}_{xz} + \mathbb{V}_{xyz}}{\mathbb{V}}$$

Interaction:

$$T_x - S_x$$



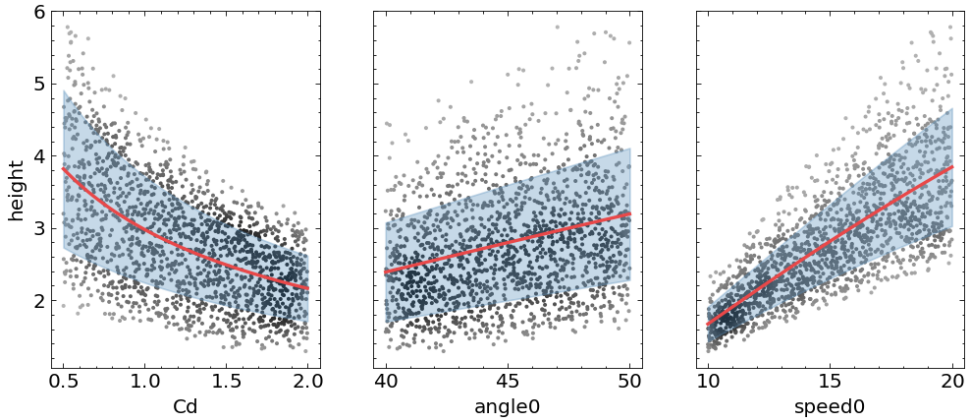




Main Effects: Connection to Model Response & Scatterplots

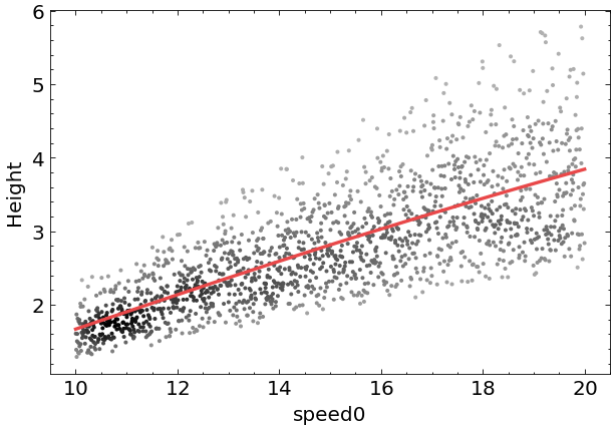
Qualitative visualization of the model behavior over its parameter domain

- Identify trend and spread.
  - *Changes* in trend suggest sensitivity
  - *Changes* in spread suggest interaction



Abscissa may need to be rank scaled. See later slides.

The trend *is* the conditional expectation



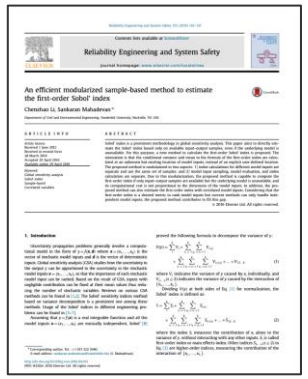
Directly related to Sobol' Index

$$S_{speed0} = \frac{\mathbb{V}\left[\mathbb{E}\left[Height|\boldsymbol{x}_{\boldsymbol{speed0}}\right]\right]}{\mathbb{V}[f]}$$

Empirical Sensitivity Analysis

Estimate Sobol' Main Effect Index without a surrogate or resampling. Use the data itself!

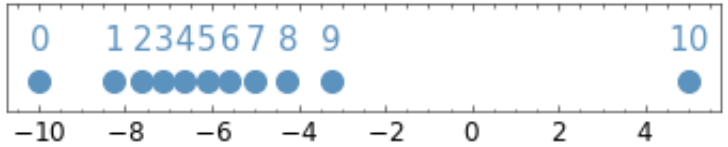
Based on [1] with some minor changes



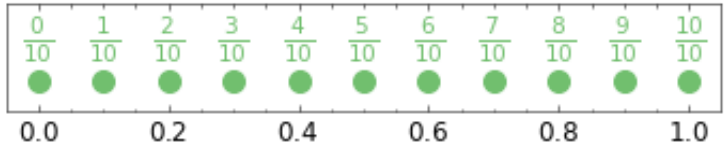
[1] C. Li and S. Mahadevan. An efficient modularized sample-based method to estimate the first-order Sobol' index. Reliability Engineering & System Safety, 153:110–121, 2016.

Step 1: Scale the data by its *rank* into [0,1]

1. Ranked from $r_i = 0, \dots, N-1$:



2. Scaled to $u_i = r_i/(N-1)$:



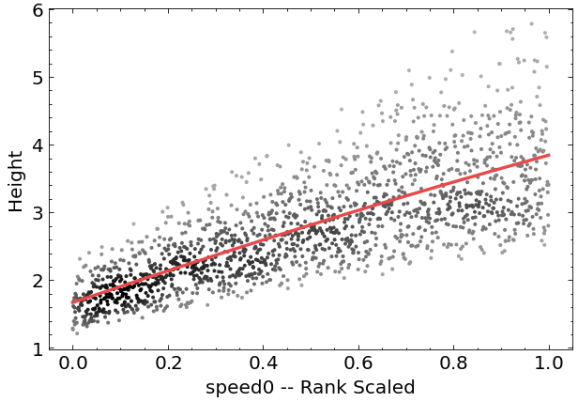
This is a simple, self-consistent, automatic, and fast scaling of the data into $\mathcal{U}[0,1]$

Step 2: Bin N samples into $n = \lfloor \sqrt{N} \rfloor$ bins

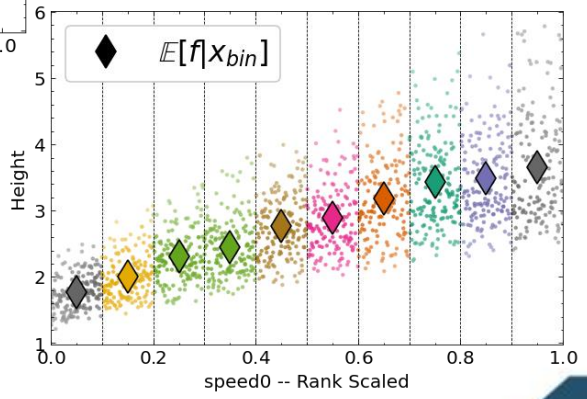
Step 3: $\mathbb{V}[\mathbb{E}[f|\alpha]] \approx \mathbb{V}[\mathbb{E}[f_i|i \in \alpha_{bin}]]$

Alternative: (better numerical performance)

$$\mathbb{V}[\mathbb{E}[f|\alpha]] = \mathbb{V}[f] - \mathbb{E}[\mathbb{V}[f|\alpha]]$$
$$\approx \mathbb{V}[f] - \mathbb{E}[\mathbb{V}[f_i|i \in \alpha_{bin}]]$$



Can also handle dependent data*



*Dependence breaks some fundamental assumptions of Sobol' Decomposition but still can provide useful results.

Example: Ocean Circulation Model Sensitivity with Time.
Hurricane Ivan, 2004

Sensitivity with time. Note that it is normalized *at each time.* Can also look at variance contribution without normalizing:

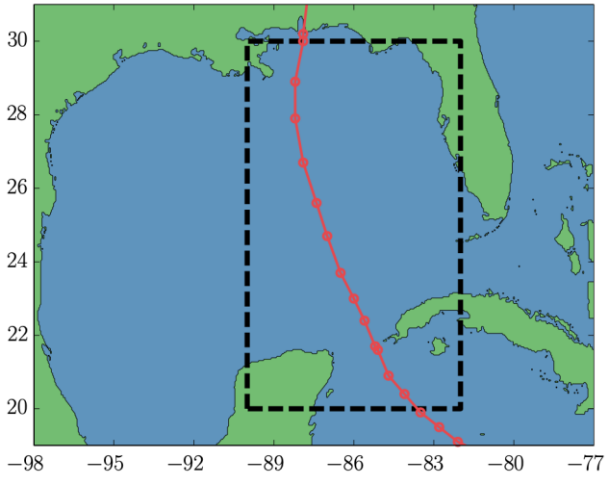
$$S_d(t) = \frac{\mathbb{V}\left[\mathbb{E}[f(t)|d]\right]}{\mathbb{V}[f(t)]} \quad \text{OR} \quad \mathbb{V}_d(t) = \mathbb{V}\left[\mathbb{E}[f(t)|d]\right]$$

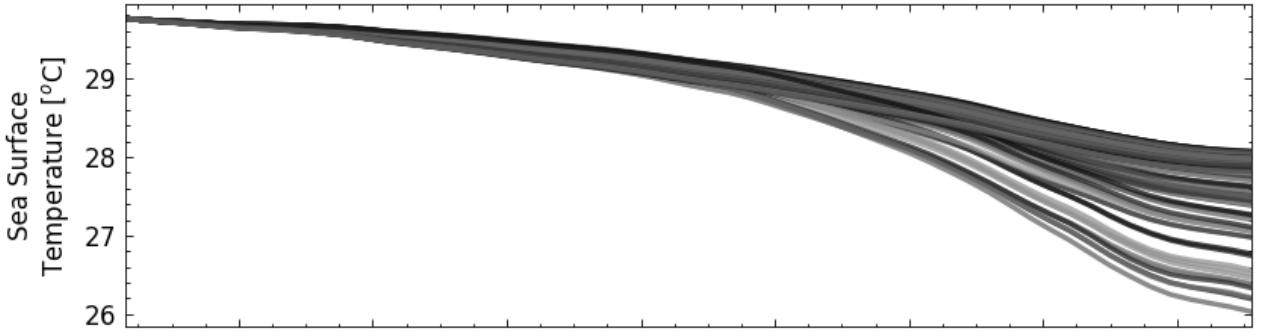
Looking *only* at Main Sensitivity (middle) tells a very different story than looking at variance (bottom).

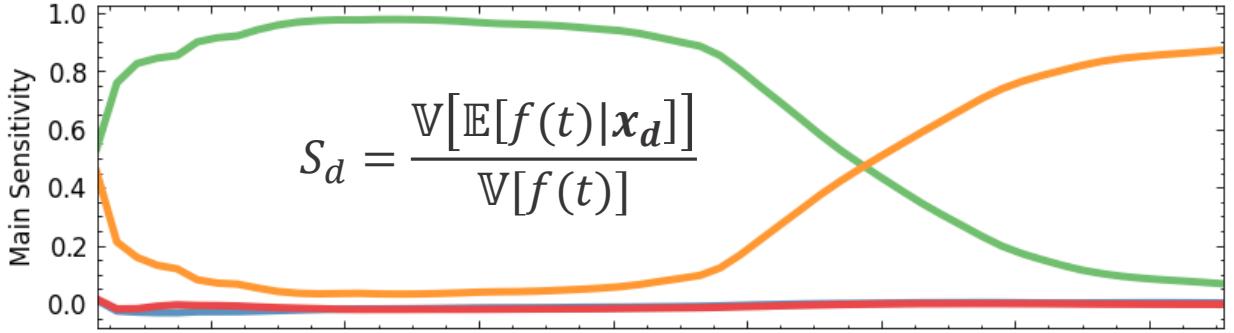
- Variance due to Background Diffusivity remained relatively constant at later times even though Main Effect went down.
- Variance from Wind Drag Coefficient (and total variance) grew in later times.

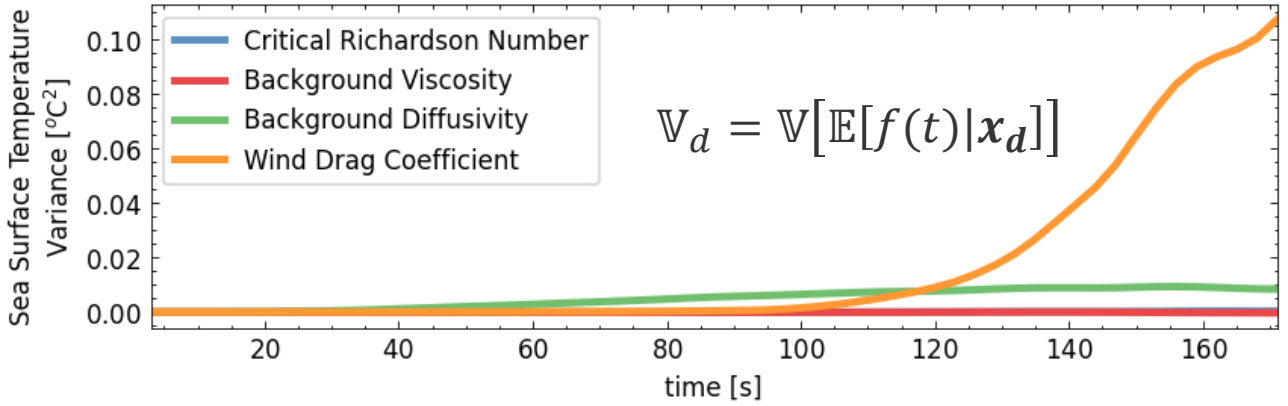
Sensitivity change corresponds to Hurricane Ivan entering the simulation domain.

Be careful of units! Do *not* take root of each variance contribution.









Legacy Data: Marthe Dataset (Strontium-90 transport in aquifer)

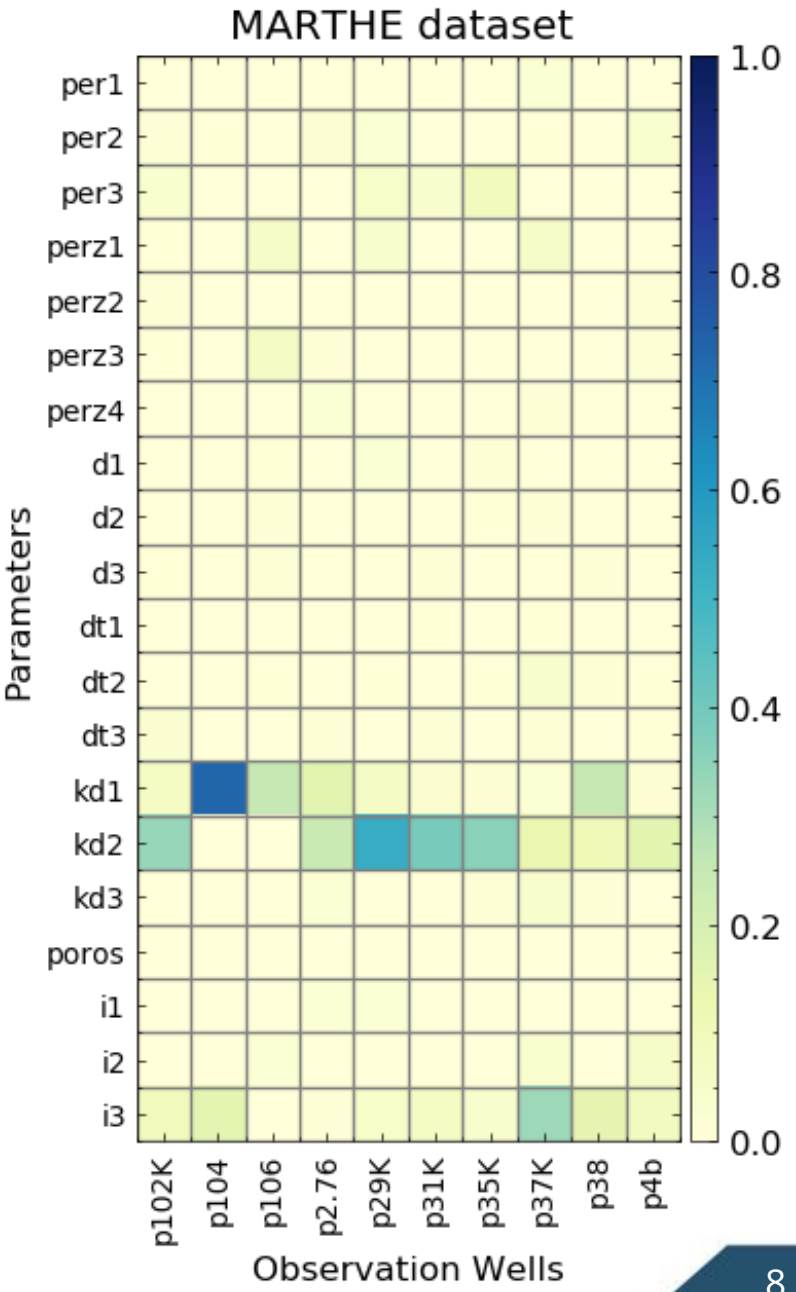
Quickly visualize sensitivity across multiple responses

Legacy Data:

- 20 input parameters
  - 5 dependent ones
- Reponses at 10 wells.
- 300 simulations

Parameter	Distribution	Description
per1	U[1, 15]	hydraulic conductivity layer 1
per2	U[5, 20]	hydraulic conductivity layer 2
per3	U[1, 15]	hydraulic conductivity layer 3
perz1	U[1, 15]	hydraulic conductivity zone 1
perz2	U[1, 15]	hydraulic conductivity zone 2
perz3	U[1, 15]	hydraulic conductivity zone 3
perz4	U[1, 15]	hydraulic conductivity zone 4
d1	U[0.05, 2]	longitudinal dispersivity layer 1
d2	U[0.05, 2]	longitudinal dispersivity layer 2
d3	U[0.05, 2]	longitudinal dispersivity layer 3
dt1	U[0.01d1, 0.1d1]	transversal dispersivity layer 1
dt2	U[0.01d2, 0.1d2]	transversal dispersivity layer 2
dt3	U[0.01d3, 0.1d3]	transversal dispersivity layer 3
kd1	W(α=1.1597, ß=19.9875)	volumetric distribution coefficient 1.1
kd2	W(α=0.891597, ß=24.4455)	volumetric distribution coefficient 1.2
kd3	W(α=1.27363, ß=22.4986)	volumetric distribution coefficient 1.3
poros	U[0.3, 0.37]	porosity
i1	U[0, 0.0001]	infiltration type 1
i2	U[i1, 0.01]	infiltration type 2
i3	U[i2, 0.1]	infiltration type 3

Weibull: $\mathrm{W}{\sim}\alpha\beta^{-\alpha}x^{\alpha-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$



1. Benchmark Proposals of GdR MASCOT-NUM. Retrieved April 2014, from http://www.gdr-mascotnum.fr/benchmarks.html.
2. Marrel, A., Iooss, B., Van Dorpe, F., & Volkova, E. (2008). An efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis, 52*, 4731-4744.
3. Volkova, E., Iooss, B., & Van Dorpe, F. (2008). Global sensitivity analysis for a numerical model of radionuclide migration from the "RRC" Kurchatov Institute radwaste disposal site. *Stochastic Environmental Research and Risk Assessment, 22*, 17-31.
4. S. Surjanovic and D. Bingham. Virtual Library of Simulation Experiments: Test Functions and Datasets. https://www.sfu.ca/~ssurjano/marthe.html

Extensions: Higher Order Sensitivity via Modified K-Means Clustering

Higher order (second order, total order) require higher-dimensional binning.

Curse of Dimensionality quickly limits utility. Even worse for dependent samples.

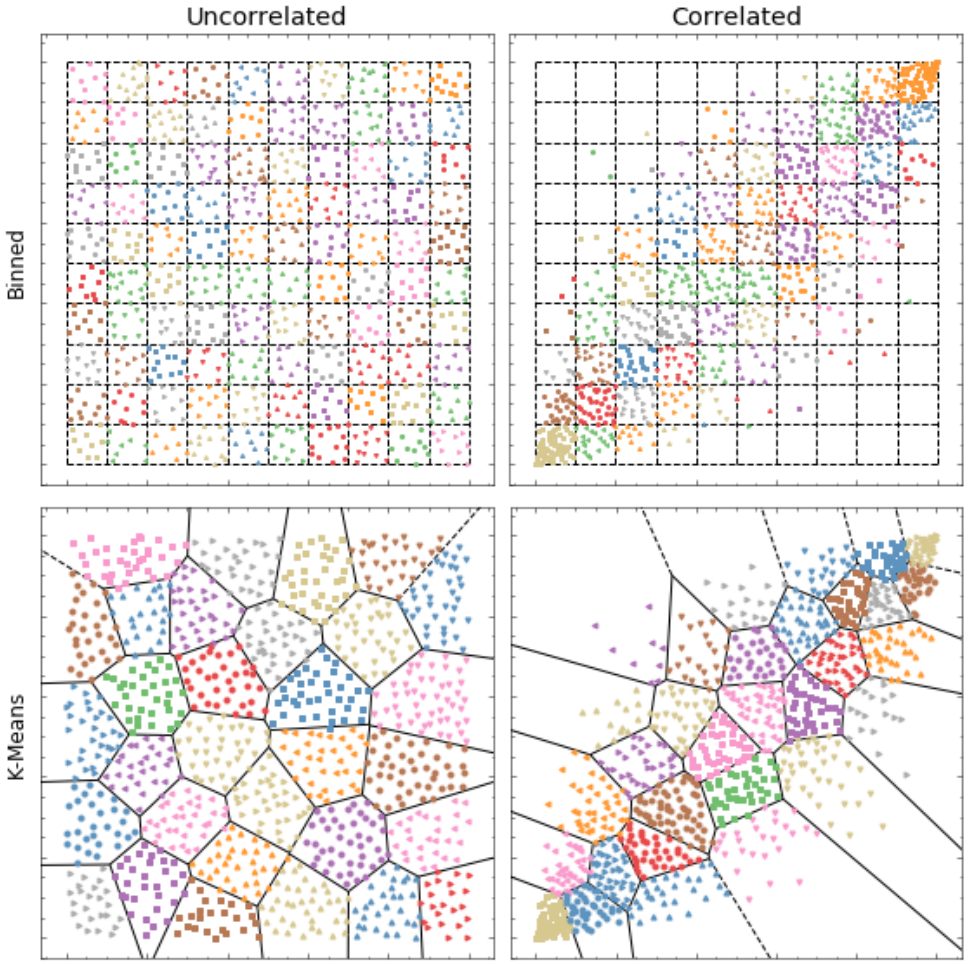
Use K-Means Clustering to *fix* the number of "bins": $n_b \approx \lfloor \sqrt[2]{N} \rceil$

- Rescale by EDF into U[0,1] unit hypercube
- Apply K-Means clustering with $\mathrm{K} = \lfloor \sqrt[2]{N} \rceil$
  - Extension: Remove outliers and/or equal size K-Means
- Compute: $\mathbb{E}[f|\boldsymbol{\alpha}] \approx \mathbb{E}[f_i|i \in \boldsymbol{\alpha}_{cluster}]$

Samples per bin are large but this reduces curse of dimensionality

Large number of conditional dimensions still impedes accuracy.

Can (still) handle dependent samples.



Extension: Error Estimation

For each N, do 1000 random trials and examine the distribution of S_d

$$S_d = \frac{\mathbb{V}[\mathbb{E}[f|x_d]]}{\mathbb{V}[f]} = 1 - \frac{\mathbb{E}[\mathbb{V}[f|x_d]]}{\mathbb{V}[f]}$$

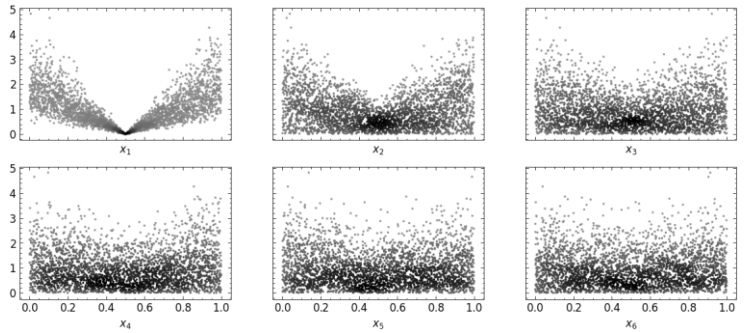
How can we estimate this uncertainty without resampling?

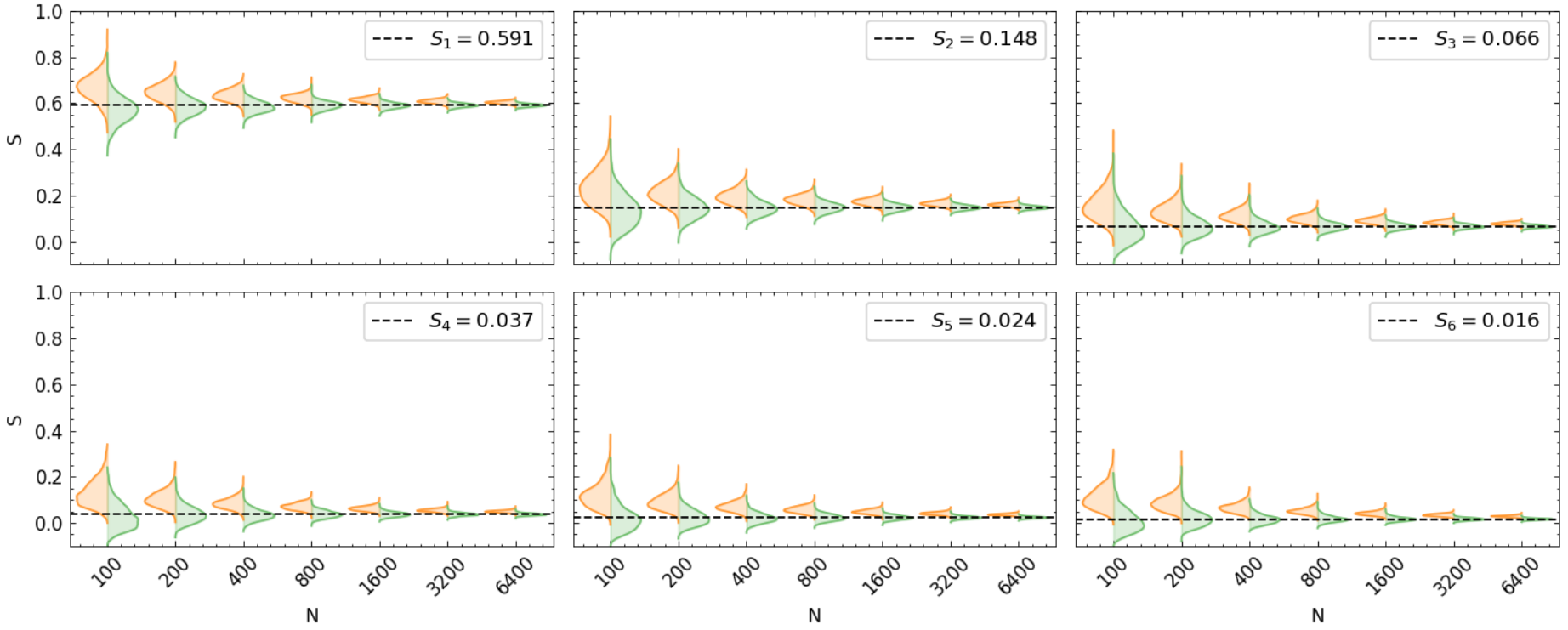
Examine statistical techniques like **bootstrap** and **jackknife**

- Construction form adds complexity and difficulty

Sobol G Function with With $a_i = i - 1$ and $\mathrm{n} = 6$

$$f(\boldsymbol{x}) = \prod_{i=1}^{n} \frac{|4x_i - 2| + a_i}{a_i + 1}$$





Extension: Greedy Sensitivity Accumulation

Total Sensitivity Indices provide an idea of sensitivity but are (a) more difficult to compute empirically, and (b) do not specify *which* parameters are interacting.

Adaptively add groups until most variance is captured. Usually $\ll 2^n$. Example: $\mathbb{V}_{\mathrm{TOL}} = 0.9$

#	Parameters				Variance Fraction
	x1	x2	x3	x4	
1	1	0	0	0	0.039
	0	1	0	0	0.281
	0	0	1	0	0.056
	0	0	0	1	0.055
Accumulated					0.431
Remaining					0.569

#	Parameters				Variance Fraction
	x1	x2	x3	x4	
1	1	0	0	0	0.039
	0	1	0	0	0.281
	0	0	1	0	0.056
	0	0	0	1	0.055
2	1	1	0	0	0.003
	0	1	1	0	0.093
	0	1	0	1	0.043
Accumulated					0.570
Remaining					0.430

#	Parameters				Variance Fraction
	x1	x2	x3	x4	
1	1	0	0	0	0.039
	0	1	0	0	0.281
	0	0	1	0	0.056
	0	0	0	1	0.055
2	1	1	0	0	0.003
	0	1	1	0	0.093
	0	1	0	1	0.043
3	1	1	1	0	0.049
	0	1	1	1	0.012
Accumulated					0.631
Remaining					0.369

#	Parameters				Variance Fraction
	x1	x2	x3	x4	
1	1	0	0	0	0.039
	0	1	0	0	0.281
	0	0	1	0	0.056
	0	0	0	1	0.055
2	1	1	0	0	0.003
	0	1	1	0	0.093
	0	1	0	1	0.043
3	1	1	1	0	0.049
	0	1	1	1	0.012
4	1	0	1	0	0.125
	0	0	1	1	0.059
Accumulated					0.815
Remaining					0.185

#	Parameters				Variance Fraction
	x1	x2	x3	x4	
1	1	0	0	0	0.039
	0	1	0	0	0.281
	0	0	1	0	0.056
	0	0	0	1	0.055
2	1	1	0	0	0.003
	0	1	1	0	0.093
	0	1	0	1	0.043
3	1	1	1	0	0.049
	0	1	1	1	0.012
4	1	0	1	0	0.125
	0	0	1	1	0.059
5	1	0	1	1	0.087
Accumulated					0.902
Remaining					0.098

Algorithm:

- Active Set $\mathcal{A}$ and Old Set $\mathcal{O}$
- Compute $\mathbb{V}_{\{x_i\}}$ for all $i \in \{1, \ldots, n_d\}$. Add all $\{x_i\}$ to $\mathcal{A}$
- While exit criteria not met:
  - Choose $\{\boldsymbol{\alpha}^*\}$ from $\mathcal{A}$ with largest $\mathbb{V}_{\{\alpha\}}$
  - Move $\{\boldsymbol{\alpha}^*\}$ from $\mathcal{A}$ to $\mathcal{O}$
  - Add $\{\boldsymbol{\alpha}'\}$ to $\mathcal{A}$ for one additional interactions with $\{\boldsymbol{\alpha}^*\}$ not in $\mathcal{A} \cup \mathcal{O}$ and compute $\mathbb{V}_{\{\boldsymbol{\alpha}'\}}$
- Stopping Criteria:
  - Capture: $\sum_{\{\boldsymbol{\alpha}\} \in \mathcal{A} \cup \mathcal{O}} \mathbb{V}_{\{\boldsymbol{\alpha}\}} \geq \mathbb{V}_{\mathrm{TOL}}$
  - Contribution: $\mathbb{V}_{\{\boldsymbol{\alpha}^*\}} \leq \mathbb{V}_{\{\boldsymbol{\alpha}^*\},TOL}$

Provides detailed, specific, information on *which* parameters are interacting. Usually does not require high-order groups

Conclusion & Summary

Assess global sensitivity through Sobol' Main Effect, i.e. fraction of the variance due *solely* to a given parameter

Use Empirical Sobol' Sensitivity to compute indices with existing and/or limited data. Minimal assumptions needed

Extensions and Open Questions

- Error estimation
- Higher order indices
- Identification of important interactions