Sobol' Indices for Sensitivity Analysis with Dependent Inputs

Joey Hart[1]

with Pierre Gremaud[1]

[1]North Carolina State University

April 16, 2018

Funding provided by NSF grants DMS-1522765 and DMS-1127914.

Formulation

Consider a function,

$$f : \mathbb{R}^p \to \mathbb{R}$$
$$\mathbf{x} \mapsto f(\mathbf{x})$$

Assume,

- $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ has some known probability distribution
- $f(\mathbf{x})$ is square integrable

Objective: determine the sensitivity of f to $\mathbf{x}$

Hierarchical Decomposition of f

$f : \mathbb{R}^p \to \mathbb{R}$ may be decomposed as

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1 \leq i < j \leq p} f_{i,j}(x_i, x_j) + \cdots + f_{1,2,\ldots,p}(x_1, x_2, \ldots, x_p)$$

Hierarchical Decomposition of f

$f : \mathbb{R}^p \to \mathbb{R}$ may be decomposed as

$$\begin{aligned} f(\mathbf{x}) &= f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1 \le i < j \le p} f_{i,j}(x_i, x_j) + \cdots + f_{1,2,\dots,p}(x_1, x_2, \dots, x_p) \\ &= f_0 + \sum_{k=1}^{p} \sum_{|u|=k} f_u(\mathbf{x}_u) \end{aligned}$$

Hierarchical Decomposition of f

$f : \mathbb{R}^p \to \mathbb{R}$ may be decomposed as

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1 \le i < j \le p} f_{i,j}(x_i, x_j) + \cdots + f_{1,2,\ldots,p}(x_1, x_2, \ldots, x_p)$$
$$= f_0 + \sum_{k=1}^{p} \sum_{|u|=k} f_u(\mathbf{x}_u)$$

where

$$f_0 = \mathbb{E}[f(\mathbf{x})]$$
$$f_i(x_i) = \mathbb{E}[f(\mathbf{x})|x_i] - f_0$$
$$f_{i,j}(x_i, x_j) = \mathbb{E}[f(\mathbf{x})|x_i, x_j] - f_i(x_i) - f_j(x_j) - f_0$$
$$\vdots$$

ANOVA (Analysis of Variance) Decomposition

$$f(\mathbf{x}) = f_0 + \sum_{k=1}^{p} \sum_{|u|=k} f_u(\mathbf{x}_u)$$

- this hierarchical decomposition exists $\forall f$ such that $f(\mathbf{x}) \in L^2$

[1]Sensitivity estimates for non linear mathematical models. Sobol'. 1993

ANOVA (Analysis of Variance) Decomposition

$$f(\mathbf{x}) = f_0 + \sum_{k=1}^{p} \sum_{|u|=k} f_u(\mathbf{x}_u)$$

- this hierarchical decomposition exists $\forall f$ such that $f(\mathbf{x}) \in L^2$
- if $x_1, x_2, \ldots, x_p$ are statistically independent then [1]

$$\mathbb{E}[f_u(\mathbf{x}_u) f_v(\mathbf{x}_v)] = 0 \quad u \neq v$$

$$\implies \text{Var}(f(\mathbf{x})) = \sum_{k=1}^{p} \sum_{|u|=k} \text{Var}(f_u(\mathbf{x}_u))$$

[1] Sensitivity estimates for non linear mathematical models. Sobol'. 1993

Sobol' Indices with Independent Variables

Using the decomposition of variance

$$\text{Var}(f(\mathbf{x})) = \sum_{k=1}^{p} \sum_{|u|=k} \text{Var}(f_u(\mathbf{x}_u)),$$

define the Sobol' indices as [2]

$$S_u = \frac{\text{Var}(f_u(\mathbf{x}_u))}{\text{Var}(f(\mathbf{x}))}.$$

- S_u measures the relative contribution of $\mathbf{x}_u$ to $\text{Var}(f(\mathbf{x}))$
- there are $2^p - 1$ Sobol' indices

[2] Sensitivity estimates for non linear mathematical models. Sobol'. 1993

First Order and Total Sobol' Indices

Define the *total Sobol' index* as

$$T_u = \sum_{v \cap u \neq \emptyset} S_v$$

In practice one frequently considers the indices

$$\{S_k, T_k\}_{k=1}^{p}$$

and refers to them as the first order and total Sobol' indices.

Interpretation of First Order and Total Sobol' Indices

$$\{S_k, T_k\}_{k=1}^{p}$$

- $S_k, T_k \in [0,1]$ measure the importance of x_k
- S_k only measures contribution of x_k
- T_k measures the contribution of all interactions involving x_k
- $S_k \leq T_k \ \forall k \in \{1,2,\ldots,p\}$
- $T_u \leq \sum\limits_{k \in u} T_k \ \forall u \subset \{1,2,\ldots,p\}$

Generalization Proposed in the Literature

$$f(\mathbf{x}) = f_0 + \sum_{k=1}^{p} \sum_{|u|=k} f_u(\mathbf{x}_u)$$

Generalization Proposed in the Literature

Taking the covariance of $f(\mathbf{x})$ with each side of

$$f(\mathbf{x}) = f_0 + \sum_{k=1}^{p} \sum_{|u|=k} f_u(\mathbf{x}_u)$$

yields

$$\text{Var}(f(\mathbf{x})) = \sum_{k=1}^{p} \sum_{|u|=k} \text{Cov}(f_u(\mathbf{x}_u), f(\mathbf{x}))$$

and gives Sobol' indices[3]

$$S_u = \frac{\text{Cov}(f_u(\mathbf{x}_u), f(\mathbf{x}))}{\text{Var}(f(\mathbf{x}))}$$

[3] Global sensitivity analysis for systems with independent and/or correlated inputs. Rabitz et al. 2010

Generalization Proposed in the Literature

$$S_u = \frac{\text{Cov}(f_u(\mathbf{x}_u), f(\mathbf{x}))}{\text{Var}(f(\mathbf{x}))}$$

The total Sobol' indices may also be generalized,

$$T_u = \sum_{v \cap u \neq \emptyset} S_v$$

Pros

- decomposes $\text{Var}(f(\mathbf{x}))$ as in the case with independence
- relates to the function decomposition

Cons

- S_u may take negative values and is no longer clear to interpret
- $S_k \not\leq T_k$ in general
- $T_u \not\leq \sum_{k \in u} T_k$

Generalization Proposed in the Literature

$$S_u = \frac{\text{Cov}(f_u(\mathbf{x}_u), f(\mathbf{x}))}{\text{Var}(f(\mathbf{x}))}$$

The total Sobol' indices may also be generalized,

$$T_u = \sum_{v \cap u \neq \emptyset} S_v$$

There are other alternative perspectives, see [4] [5]

[4] Estimation of global sensitivity indices for models with dependent variables. Kucherenko et al. 2012.

[5] Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Mara et al. 2015.

A Simple Example to Illustrate

$$f(\mathbf{x}) = 20x_1 + 16x_2 + 12x_3 + 4x_4$$

$$\mathbf{x} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .5 & .5 & .8 \\ .5 & 1 & 0 & 0 \\ .5 & 0 & 1 & .3 \\ .8 & 0 & .3 & 1 \end{bmatrix} \right)$$

S_1	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	$S_{1,2,3}$	$S_{1,2,4}$	$S_{1,3,4}$	$S_{1,2,3,4}$	T_1
.903	-.393	-.333	-.295	.030	-.047	.154	-.010	.009

Table: Sobol' indices involving variable x_1.

A Simple Example to Illustrate

$$f(\mathbf{x}) = 20x_1 + 16x_2 + 12x_3 + 4x_4$$

$$\mathbf{x} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .5 & .5 & .8 \\ .5 & 1 & 0 & 0 \\ .5 & 0 & 1 & .3 \\ .8 & 0 & .3 & 1 \end{bmatrix} \right)$$

T_1	T_2	$T_{1,2}$
.009	.021	.450

Table: Total Sobol' indices.

$$T_{1,2} \not\leq T_1 + T_2$$

Approximation Theoretic Perspective

- let $f_0 = \mathbb{E}[f(\mathbf{x})]$
- let $u \subset \{1, 2, \dots, p\}$ and $\sim u = \{1, 2, \dots, p\} \setminus u$
- let $\mathcal{P}_{\sim u} f$ be the optimal L^2 approximation of $f - f_0$ in the space of functions which do not depend on $\mathbf{x}_u$

Theorem: $T_u = \frac{||(f - f_0) - \mathcal{P}_{\sim u} f||_2^2}{||f - f_0||_2^2}$

Interpretation: $T_u \in [0, 1]$ measures the squared relative error of approximating $f - f_0$ by a function which does not depend on $\mathbf{x}_u$.

Uses of Global Sensitivity Analysis

Global sensitivity analysis may be used to:

- prioritize data acquisition and/or model development
- determine unimportant variables for dimension reduction
  - guide the construction of surrogate models
  - "freezing" variables
- determine and mitigate risk
- gain insight into underlying phenomenon

Uses of Global Sensitivity Analysis

Global sensitivity analysis may be used to:

- prioritize data acquisition and/or model development
- determine unimportant variables for dimension reduction
  - guide the construction of surrogate models
  - "freezing" variables
- determine and mitigate risk
- gain insight into underlying phenomenon

The remainder of this talk will focus on "freezing" (replacing) unimportant variables to reduce dimensions.

GSA for Dimension Reduction

- If T_u is small and I freeze $\mathbf{x}_u$, how much error do I incur?
- Results exist when the variables are independent. [6]
- Dependencies among the variables help.
- Replacing $\mathbf{x}_u$ with $g(\mathbf{x}_{\sim u})$ is superior to freezing.
- T_u is the error when $f(\mathbf{x})$ is projected.
- The approximation theoretic perspective also helps analyze the error from replacing $\mathbf{x}_u$.

[6] Estimating the approximation error when fixing unessential factors in global sensitivity analysis. Sobol' et al. 2007.

Replacing Unimportant Variables

$$f(\mathbf{x}) = f_0 + \mathcal{P}_{\sim u} f(\mathbf{x}_{\sim u}) + \mathcal{P}_{\sim u}^{\perp} f(\mathbf{x})$$

$$T_u = \frac{||\mathcal{P}_{\sim u}^{\perp} f(\mathbf{x})||_2^2}{||f(\mathbf{x}) - f_0||_2^2}$$

$$\delta_u = \frac{||f(\mathbf{x}) - f(g(\mathbf{x}_{\sim u}), \mathbf{x}_{\sim u})||_2^2}{||f(\mathbf{x}) - f_0||_2^2}$$

$$= \frac{||\mathcal{P}_{\sim u}^{\perp} f(\mathbf{x}) - \mathcal{P}_{\sim u}^{\perp} f(g(\mathbf{x}_{\sim u}), \mathbf{x}_{\sim u})||_2^2}{||f(\mathbf{x}) - f_0||_2^2}$$

Lemma

For any $u \subset \{1, 2, \ldots, p\}$ and any g, $\delta_u \geq T_u$

Replacing Unimportant Variables

$$f(\mathbf{x}) = f_0 + \mathcal{P}_{\sim u} f(\mathbf{x}_{\sim u}) + \mathcal{P}_{\sim u}^{\perp} f(\mathbf{x})$$

$$T_u = \frac{||\mathcal{P}_{\sim u}^{\perp} f(\mathbf{x})||_2^2}{||f(\mathbf{x}) - f_0||_2^2}$$

$$\delta_u = \frac{||\mathcal{P}_{\sim u}^{\perp} f(\mathbf{x}) - \mathcal{P}_{\sim u}^{\perp} f(g(\mathbf{x}_{\sim u}), \mathbf{x}_{\sim u})||_2^2}{||f(\mathbf{x}) - f_0||_2^2}$$

- $\delta_u \geq T_u$
- Would like to have an upper bound on δ_u
- Difficult to bound tightly in general

Replacing Unimportant Variables

$$f(\mathbf{x}) = f_0 + \mathcal{P}_{\sim u} f(\mathbf{x}_{\sim u}) + \mathcal{P}_{\sim u}^{\perp} f(\mathbf{x})$$

$$T_u = \frac{||\mathcal{P}_{\sim u}^{\perp} f(\mathbf{x})||_2^2}{||f(\mathbf{x}) - f_0||_2^2}$$

$$\delta_u = \frac{||\mathcal{P}_{\sim u}^{\perp} f(\mathbf{x}) - \mathcal{P}_{\sim u}^{\perp} f(g(\mathbf{x}_{\sim u}), \mathbf{x}_{\sim u})||_2^2}{||f(\mathbf{x}) - f_0||_2^2}$$

- δ_u will be small when

$$||\mathcal{P}_{\sim u}^{\perp} f(g(\mathbf{x}_{\sim u}), \mathbf{x}_{\sim u})||_2 \approx ||\mathcal{P}_{\sim u}^{\perp} f(\mathbf{x})||_2$$

- δ_u is small if T_u is robust to changes in the distribution of $\mathbf{x}$

The g-function

$$f(\mathbf{x}) = \prod_{k=1}^{10} \frac{|4x_k - 2| + a_k}{1 + a_k}$$

where $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, i.e. is normally distributed

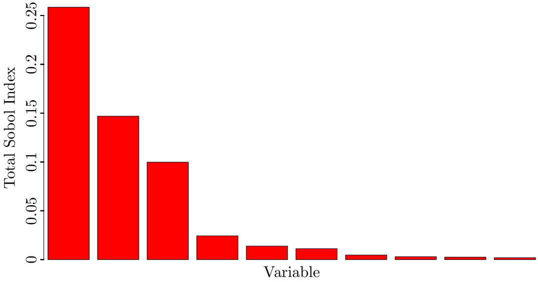
$$\mu_i = \frac{1}{2} \qquad i = 1, 2, \ldots, 10$$

$$\Sigma_{i,i} = \frac{1}{6} \qquad i = 1, 2, \ldots, 10$$

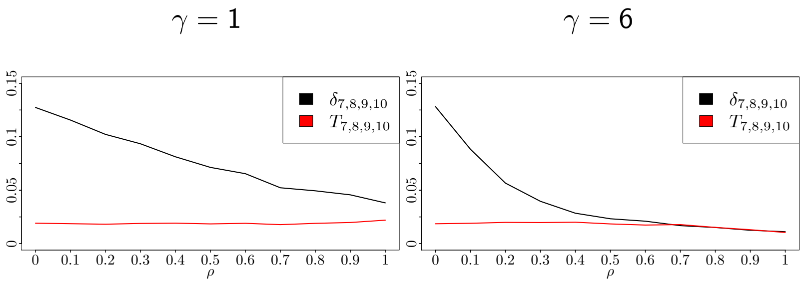
$$\Sigma_{i,j} = \frac{\rho}{6|i-j+1|^{\frac{1}{\gamma}}} \qquad i \neq j$$

- larger $\rho \implies$ stronger correlations
- larger $\gamma \implies$ "dense" covariance matrix

Total Sobol' Indices with $\rho = \frac{1}{2}$ and $\gamma = 6$



Approximation Errors as ρ Varies ($g(\mathbf{x}_{\sim u}) = \mathbb{E}[\mathbf{x}_u | \mathbf{x}_{\sim u}]$)



- stronger correlations $\implies$ δ_u decreases
- $\delta_u \to T_u$ faster when the correlations are "dense"

Summary

- Reinterpreted the total Sobol' index T_u in terms of approximation error instead of variance analysis.
- Gives T_u a clear interpretation in terms of optimal approximation errors.
- Provides a framework to analyze the error when replacing unimportant variables.
- Argue that dependencies can help reduce error.

Some Questions

- Test for the robustness of T_u with respect to changes in the distribution of $\mathbf{x}$? Bound δ_u using this?
- Approaches for constructing $g(\mathbf{x}_{\sim u})$ when $\mathbf{x}_u$ and $\mathbf{x}_{\sim u}$ have nonlinear dependencies? How does this relate to robustness of the Sobol' indices?
- Other characterizations of Sobol' indices which are more useful for other applications of global sensitivity analysis? Maybe involving other methods (moment-independent importance measures, Shapley values, derivative-based global sensitivity measures).

Questions?

Joey Hart

North Carolina State University

jlhart3@ncsu.edu

J. Hart and P.A. Gremaud. An approximation theoretic perspective of Sobol' indices with dependent variables.
https://arxiv.org/pdf/1801.01359v2.pdf

J. Hart and P.A. Gremaud. Robustness of the Sobol' indices to distributional uncertainty.
https://arxiv.org/pdf/1803.11249.pdf