Advances and challenges in global sensitivity analysis

Pierre Gremaud (NC State)

joint work with:

- Joey Hart (Sandia NL)
- Tim David (U. of Canterbury, New Zealand)
- Alen Alexanderian, Mike Merritt, Robin Morillo, Ralph Smith (NC State)
but first: Happy 728th Birthday, Switzerland (Aug. 1, 1291)
what I usually do on Aug. 1
model: \( q = g(\theta_1, \ldots, \theta_p) \)

typically

- \( g \) is a computer code
- parameters are **uncertain**
- \( p \) is large
$p$ large?

- **in stat speak**: $p$ and $n$ (sample size/data) matters
  - classic stat: $p$ is fixed, $n \to \infty$,
  - law of large numbers, central limit theory
  - high dim: $p$ and $n$ large, often with $p \gg n$

- **in math speak**: standard approaches do not work
  - objects can’t be handled $\Rightarrow$ sampling
  - low precision arithmetics(!)
  - see the OTHER RTG: randomized numerical analysis
sensitivity analysis \((q = g(\theta))\)

we want to:

*quantify how uncertainties in the model response can be apportioned to uncertainties in model inputs*

the larger the contribution, the more important the input
rationale for SA (inspired by Saltelli)

- model corroboration: is the inference robust?
- research prioritization: which factor most deserves further analysis/measurement?
- model simplification: can factors/compartments be fixed or simplified?
- model reliability: identify factors which interact and may lead to extreme values
GSA challenges

- no agreement on the meaning of important
- one SA method $\iff$ one definition of "importance"
- inputs can be correlated
- GSA results depend on how parameter uncertainty is modeled; robustness?
- meaning of GSA for evolution pbs; causality?
- practical considerations $\Rightarrow$ use of surrogates (often):
  - surrogate $\approx$ model $\Rightarrow$ GSA(surrogate) $\approx$ GSA(model)
importance?

\[ g(\theta_1, \theta_2) = \sin^2 \beta \theta_1 \sin^2 \theta_2 \]  \[ \theta_i \sim U(0, 2\pi), \ i = 1, 2 \]

- dashed lines: partial derivative importance
- solid lines: total Sobol’ indices
- only agree for \( \beta = 1 \!)
GSA: lots of choices

- regression based
- variance based (Sobol’ indices)
- derivative based (Morris screening, ...)
- game theoretic (Shapley values/effects)
- and others...

this talk: (mostly) variance based
variance based GSA

Ilya Sobol’

- considers $\theta_i$’s as random variables
- apportions to them their relative contribution to the variance of the response

trivial example: $q = a\theta_1 + b\theta_2$, $\theta_i \sim N(0, \sigma_i^2)$, $a, b > 0$

$q \sim N(0, \sigma_q^2)$ with $\sigma_q^2 = a^2\sigma_1^2 + b^2\sigma_2^2$

$$
\Rightarrow 1 = \frac{a^2\sigma_1^2}{a^2\sigma_1^2 + b^2\sigma_2^2} + \frac{b^2\sigma_2^2}{a^2\sigma_1^2 + b^2\sigma_2^2}
$$

- note the importance of the $\sigma_i$’s!
total Sobol’ indices

law of total variance

\[ \text{var}(\mathbb{E}[q|\theta_{\sim i}]) + \mathbb{E}[\text{var}(q|\theta_{\sim i})] = \text{var}(q) \]

and thus

\[ \text{var}(q) - \text{var}(\mathbb{E}[q|\theta_{\sim i}]) = \mathbb{E}[\text{var}(q|\theta_{\sim i})] \]

remaining variance if \( \theta_{\sim i} \) were known

- total index: \( T_i = \frac{\mathbb{E}[\text{var}(q|\theta_{\sim i})]}{\text{var}(q)} = 1 - \frac{\text{var}(\mathbb{E}[q|\theta_{\sim i}])}{\text{var}(q)} \)
\( T_i = 0 \iff \theta_i \text{ non-important} \)

\( \iff \)

\( \theta_i \text{ non-import.} \implies \var(q|\theta \sim i) = 0 \implies \mathbb{E}[\var(q|\theta \sim i)] = 0 \implies T_i = 0 \)

\( \implies \)

\( T_i = 0 \implies \mathbb{E}[\var(q|\theta \sim i)] = 0 \implies \var(q|\theta \sim i) = 0 \implies \theta_i \text{ not import.} \)

Unimportance is important! (Art Owen)

- allows focus on key inputs
- potential for faster codes
ANOVA (Reader’s Digest version)

- assume $\theta_i$, iid, $\theta_i \sim U(0, 1)$
- split $\theta = (\theta_i, \theta_{\sim i})$ and decompose $g$ as

\[ g(\theta) = g_0 + g_1(\theta_i) + g_2(\theta_{\sim i}) + g_{12}(\theta_i, \theta_{\sim i}) \]

where

- $g_0 = \int g(\theta) \, d\theta$,
- $g_1(\theta_i) = \int (g - g_0) \, d\theta_{\sim i}$, $g_2(\theta_{\sim i}) = \int (g - g_0) \, d\theta_i$
- $g_{12}$ = remainder

- above functions have zero average $\Rightarrow \perp \Rightarrow$

\[
\text{var}(q) = \int (g(\theta) - g_0)^2 \, d\theta = \int g(\theta)^2 \, d\theta - g_0^2
\]

\[
= \int g_1^2 \, d\theta + \int g_2^2 \, d\theta + \int g_{12}^2 \, d\theta
\]

\[
\underline{\text{var}(g_1)} + \underline{\text{var}(g_2)} + \underline{\text{var}(g_{12})}
\]
another way to look at things

**equivalent definition:**

\[
T_i = \frac{\var_i}{\var(q)}; \quad S_i = \frac{\var(g_i)}{\var(q)}
\]

- **total index**
- **1st order index**

where

\[
\var_i = \var(g_1) + \var(g_{12}) = \text{total variance corresponding to } \theta_i
\]

**exercise:**

\[
\var_i = \frac{1}{2} \iint \left( \frac{\partial g}{\partial \theta_i}(\hat{\theta})(\theta_i - \theta_i') \right) d\theta d\theta_i'
\]

where \(\theta' = (\theta_1, \ldots, \theta_{i-1}, \theta_i', \theta_{i+1}, \ldots, \theta_p)\).
independent variables: summary

notation: \( u \subset \{1, \ldots, p\} \)

\[
\sum_{k=1}^{p} \sum_{|u|=k} S_u = 1 \quad \text{conservation}
\]

\( \forall u, 0 \leq S_u \leq 1 \quad \text{boundedness} \)

\[
\min_{k \in u} T_k \leq T_u \leq \sum_{k \in u} T_k
\]

\( S_u \) close to 1 \( \Rightarrow \) \( \theta_u \) is important

\( T_u \) close to 0 \( \Leftrightarrow \) \( \theta_u \) is non-important
dependent variables?\textsuperscript{1}

\[ S_u = \frac{\text{cov}(g_u, g)}{\text{var}(q)} , \quad T_u = \sum_{v \cap u \neq \emptyset} S_v \]

\textsuperscript{1}Chastaing et al., J. Stat. Comput. Sim., 2012
however...

\[ \sum_{k=1}^{p} \sum_{|u|=k} S_u = 1 \quad \text{conservation} \quad \text{YES} \]

\[ \forall u, 0 \leq S_u \leq 1 \quad \text{boundedness} \quad \text{NO} \]

\[ \min_{k \in u} T_k \leq T_u \leq \sum_{k \in u} T_k \quad \text{NO} \]

\[ S_u \, \text{close to 1} \Rightarrow \theta_u \, \text{is important} \quad \text{NO} \]

\[ T_u \, \text{close to 0} \Leftrightarrow \theta_u \, \text{is non-important} \quad \text{NO} \]

both interpretation and computations are challenging!
\[ g(\theta) = 20\theta_1 + 16\theta_2 + 12\theta_3 + 10\theta_4 + 4\theta_5, \quad \text{with } \theta \sim N(\mu, \Sigma) \]

\[
\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0.5\rho & 0.5\rho & 0 & 0.8\rho \\ 0.5\rho & 1 & 0 & 0 & 0 \\ 0.5\rho & 0 & 1 & 0 & 0.3\rho \\ 0 & 0 & 0 & 1 & 0 \\ 0.8\rho & 0 & 0.3\rho & 0 & 1 \end{bmatrix}, \quad 0 \leq \rho \leq 1.
\]

for \( \rho = 1 \):
\[
T_1 = 0.0087, \quad T_2 = 0.0196, \quad T_{1,2} = 0.4228
\]
Ig square integrable; $g_0 = \mathbb{E}[g(\theta)]$

- a key question:

- how accurately can $g(\theta) - g_0$ be approximated WITHOUT $\theta_u$?

  or

  what is the error associated with the approximation $g(\theta) - g_0 \approx \mathcal{P}_{\sim u}g(\theta\sim u)$?

$\mathcal{P}_{\sim u}g$: optimal $L^2$ approximation which does not depend on $\theta_u$
approximation theoretic perspective

Theorem (Hart, G., 2018)

\[ T_u = \frac{\| (g(\theta) - g_0) - \mathcal{P}_{\sim u}g(\theta_{\sim u}) \|_2^2}{\| g(\theta) - g_0 \|_2^2}. \]
dimension reduction

ways to get a "smaller" model:

- project onto subspace of fcts of fewer var. (expensive)
- train a surrogate model using fewer variables
- fix some of the $\theta_i$’s to "nominal" values (common)

above result helps: natural dimension reduction for Sobol
is the above projection:

$$T_u = 1 - \frac{\| \mathcal{P}_{\sim u} g \|^2_2}{\| g - g_0 \|^2_2}$$

ongoing work: efficient ways to compute $\mathcal{P}_{\sim u} g$ "in passing"
dimension reduction II

- $\theta = (\theta_u, \theta_{\sim u})$; assume $T_u$ small
- we want:

$$g(\theta) \approx g(h(\theta_{\sim u}), \theta_{\sim u})$$

for instance: $h(\theta_{\sim u}) = \mathbb{E}[\theta_u]$

key quantity: $\delta_u = \frac{\|g(\theta) - g(h(\theta_{\sim u}), \theta_{\sim u})\|_2^2}{\|g(\theta) - g_0\|_2^2}$

Theorem (Hart, G., 2018)

$$T_u \leq \delta_u \leq 4T_u$$
it is possible to evaluate robustness of $T_u$ at "no" extra cost

involves regarding $T_u$ as $\{PDF\} \rightarrow \{Sobol indices\}$ and taking derivatives (Fréchet differentiable) and... evaluating carefully

limitations...
time dependent GSA

\[ y'' + 2\alpha y' + (\alpha^2 + \beta^2)y = 0, \]
\[ y(0) = \ell, \quad y'(0) = 0 \]

underdamped mechanical oscil: \( y = \ell e^{-\alpha t}(\cos \beta t + \frac{\alpha}{\beta} \sin \beta t) \)
time dependent GSA II

- yardstick! $T_i(t)$ not very informative
- better way: $\mathcal{T}_i(t) = \frac{\int_0^t \text{var}_i \, d\tau}{\int_0^t \text{var} \, q \, d\tau}$. (Gamboa et al. 2014)
an application (with T. David and J. Hart)

- 4 pathways
- 67 state variables
- 160 parameters
- 4 "objects"/models
  - NE
  - AC
  - SMC, EC
  - WM
- share 13 variables (coupling)

picture from T. David et al.
experimental setup

We've increased the voltage of the electrical shocks. Now, the male mouse puts the little toilet seat down almost every time.
Goals

- **physiology**: understand *dominant* cellular mechanisms resulting in cerebral tissue perfusion after neuronal stimulation
- **diagnostics (understanding)** rather than *prognostics* (predictions)
- **mathematics**: develop GSA methods for *large* systems of the type

\[
\frac{dy}{dt} = f(\theta, y) + \text{ICs}
\]

**QoI**: \( q = G(\theta, y(\theta)) \equiv g(\theta) \)
Challenges

- large systems of ODEs: state variables $y = (y_1, \ldots, y_{67})$
- large number of uncertain parameters $\theta = (\theta_1, \ldots, \theta_{160}) \Rightarrow$ high-dimensionality

$$q = g(\theta_1, \ldots, \theta_{160})$$

- standard GSA methods are too expensive out of the box
- multiple time scales $\Rightarrow$ stiffness
- "fuzzy" goal $\Rightarrow$ several QoIs
hunting for feasible points

919 initial samples ($\theta_k$) without stimulus; of these

- 670 fail (the MATLAB ODE solver ode15s can’t take time step small enough)
- 110 reach an **unstable** steady-state and then go nuts
- 139 reach a **stable** steady-state
Pearson $\Rightarrow$ two parameters are strongly correlated
correlated pair

*: premature term.
+ : unstable steady state
○ : stable steady state
let’s sample "harder"

updated distribution:

- ~correlated pair: as before (iid. U. ±10%)
- correlated pair: fitted Frank copula with beta marginals

\[ CDF = c(F_i(\theta_i), F_j(x_j)) \]

- with \( F_i(x_i) = Pr(\theta_i \leq x_i) = \frac{B(x_i; \alpha_i, \beta_i)}{B(1, \alpha_i, \beta_j)} \) with

\[ B(x; \alpha, \beta) = \int_0^x t^{\alpha-1}(1 - t)^{\beta-1} \, dt \]

- \( c(u, v) = -\frac{1}{\mu} \log \left( 1 + \frac{\exp(-\mu u) - 1)(\exp(-\mu v) - 1)}{\exp(-\mu) - 1} \right) \)
iterative process

- while ∼ tired or ∼ converged
  - fit above bivariate distribution to $(\theta_i, \theta_j)$
  - sample \( \theta \)
  - solve
  - keep stable steady state solutions

in the end: 902 stable steady solutions
dimension reduction and GSA roadmap

for each QoI
  ▶ reduce dimension through linear regression
  ▶ fit polynomial chaos surrogate on reduced set of parameters
  ▶ compute total Sobol’ indices from PCE
linear regression

\[ QoI = g(\theta) \approx \beta_0 + \sum_{k=1}^{160} \beta_k \theta_k \]

- importance of \( k \)-th parameter

\[ L_k = \frac{|\beta_k|}{\sum_{j=1}^{160} |\beta_j|}, \quad k = 1, \ldots, 160 \]

- retain only \( \theta_k \)'s with \( L_k > 0.01 \) for nonlinear fitting
PCE+ Sobol’': flow

**Linear Regression Surrogate**

- Model Value vs. Surrogate Prediction
- Parameter Importance in Linear Regression Surrogate

**PCE Surrogate**

- Model Value vs. Surrogate Prediction
- Total Sobol Indices of PCE

![Graphs and plots](attachment:image.png)
Perspectives and Conclusions

- lots of work to do in high dim approximation
- dimension reduction is key
- how to construct surrogate models
- to solve a specific problem, quantitative experts and field experts have to work together
- there may be "cultural issues" (is everyone happy with a linear model with 10 parameters?)
papers

- J.L. Hart, P. Gremaud, *An Approximation Theoretic Perspective of Sobol’ Indices with Dependent Variables*, Int. J. Uncertainty Quant., 2018
- A. Alexanderian, P. Gremaud, R.C. Smith, *Variance-based sensitivity analysis for time-dependent processes*, under revision