Center for Science of Information
NSF Science and Technology Center

Information Theoretic Sensitivity Analysis for Nondeterministic Simulators

Yu-Jay Huoh
UC Berkeley Statistics

Introduction

In all fields of science there is growing emphasis on using computer code to model complex systems. Some examples are liquid flow through fractured media or ocean-atmosphere circulation. Computer models are useful for many things - forecasting, finding optimal settings, or developing understanding. One particular use for simulators is sensitivity analysis - studying how uncertainty in the inputs affects the output of the simulator.

One challenge that arises when working with computer models are *long running times*. Traditional analysis often requires many evaluations of the simulator to work properly. When a model takes hours or days to run, these methods can't be used and novel approaches must be taken to study these models.

More recently *nondeterministic models* have become more popular. Traditional computer simulators are deterministic - for a fixed set of inputs, they always return the same output. A *nondeterministic model* returns different values when run multiple times at the same inputs. Nondeterministic models present specific challenges when conducting sensitivity analysis because traditional methods assume any uncertainty in the output is due entirely to uncertainty in the inputs.

Sensitivity Analysis

- This poster focuses on *global sensitivity analysis* which is the study of how uncertainty in the inputs affects the output of a computer simulator.
  - Small changes in an input may cause a dramatic change in the output.
- Analysis can be on
  - One input with all the other inputs fixed - *first order analysis*
  - Groups of inputs with rest of inputs fixed - *interactions or higher order analysis*
  - An input and all of its combined effects - *total sensitivity analysis*
- Sensitivity analysis helps researchers decide problems such as
  - Which input parameters have the most influence on the output? Are there any non-influential inputs that we can safely fix?
  - How should resources for reducing input uncertainty be allocated to best decrease output variability?
- The setup:
  - f - the computer model or simulator, which is some function
  - $X_1, \ldots, X_p$ - inputs to the simulator
  - $Y = f(X_1, \ldots, X_p)$ - output of the simulator at a given set of inputs
  - Usually the simulator is run at a collection of points, and we denote each input-output pair in the collection as (X_i, Y_i) where $Y_i = f(X_{i,1}, \ldots, X_{i,p})$
- If f is deterministic, can perform SA through variance decompositions. [5]
  - If f is square integrable, can always write it as sum of orthogonal functions

$$f(X_1, \ldots, X_p) = f_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{i<j} f_{ij}(X_i, X_j) + \ldots + f_{1\ldots p}(X_1, \ldots, X_p)$$

  - This decomposition is also valid for variance of $Y = f(X_1, \ldots, X_p)$

$$Var(Y) = \sum_{i=1}^{p} Var[f_i(X_i)] + \sum_{i<j} Var[f_{ij}(X_i, X_j)] + \ldots + Var[f_{1\ldots p}(X_1, \ldots, X_p)]$$

  - The variance components $Var[f_i(X_i)]$ can interpreted as *sensitivity indices* and are sometimes called *Sobol indices*.
- Usually f is not known in closed form, so the indices must be estimated through Monte Carlo and repeated function evaluations.
- If f is a slow to evaluate function, it's not possible to compute Monte Carlo type estimates directly.
  - One solution is to use a *functional surrogate* or *response surface approximation* that is quick to evaluate in place of f
  - Estimate the indices for the approximation through repeated evaluations. If the approximation is close, the estimated indices will be close to the true values.
  - Oakley and O'Hagan used Bayesian Emulators as response surface [4]
- For a nondeterministic function, the variance can't be decomposed this way because each set of inputs gives a distribution for the output.

Information Theoretic Sensitivity Analysis

- Instead of characterizing the effect of $X_1, \ldots, X_p$ on Y through means and variances, look at the effect of the inputs on the entropy of Y
  - Useful for nondeterministic simulators because entropy characterizes distributions.
- Define the *mutual information index (MII)* for an input X_i to be [2]

$$S_i = \frac{I(Y, X_i)}{H(Y)} = \frac{H(Y) - H(Y|X_i)}{H(Y)}$$

  - $H(Y)$ is the entropy of Y
  - $H(Y|X_i)$ is the conditional entropy of Y given X_i
  - $I(Y, X_i)$ is the mutual information between Y and X_i.
- The MII is the normalized reduction in uncertainty in Y caused by knowing X_i, measured in proportion of bits.
- There are analogous expression for higher order terms (interactions) and total sensitivity indices, but these are less understood and much harder to estimate.

Traditional Estimation

- For estimation, focus on just the numerator - $I(Y, X_i)$. The denominator is only for normalization.
- Recall $H(Y|X) = H(Y, X) - H(X)$ (chain rule), so information is

$$I(Y, X_i) = H(Y) - H(Y|X_i) = H(Y) + H(X_i) - H(Y, X_i)$$

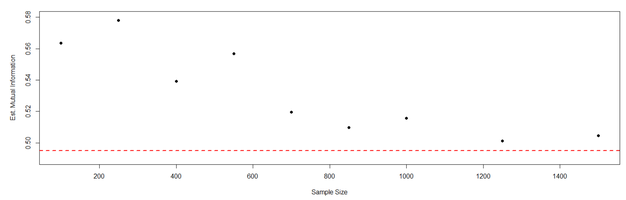
  Estimation of mutual information is equivalent to estimating three entropies.
- Given a sample $X_1, \ldots, X_n$ from some distribution F, the *resubstitution estimate* of the entropy $H(X)$, $X \sim F$ is

$$\hat{H}(X) = -\frac{1}{n} \sum_{j=1}^{n} log\hat{f}_n(X_j)$$

  where $\hat{f}_n$ is some density estimate. [1]
  - If f_n is a kernel density estimate, then $\hat{H}(X)$ is a mean square consistent estimate of $H(X)$.
  - If f_n is a histogram estimate, then $\hat{H}(X)$ is a $\sqrt{n}$ consistent estimate of $H(X)$.
- Given a collection of runs $(X_j, Y_j), j = 1, \ldots n$, we can estimate S_i by
  1. Estimate $\hat{f}(y), \hat{f}(x_i), \hat{f}(y, x_i)$ - these can be histogram or kernel density estimates.
  2. Use these density estimates to estimate the entropies $\hat{H}(Y), \hat{H}(X_i), H(\hat{Y}, X_i)$.
  3. The estimate of the MII will be

$$\hat{S}_i = \frac{\hat{H}(Y) + \hat{H}(X_i) - H(\hat{Y}, X_i)}{\hat{H}(Y)}$$

Estimation for Slow to Evaluate Nondeterministic Simulators



Simulation study - estimated mutual information against sample size.

- Approach will not work for slow to evaluate simulators - not enough data points for asymptotics like consistency
- Two approximations in the resubstitution estimate
  1. Monte Carlo integral approximation: $-\frac{1}{n}\sum_{j=1}^{n} logf(X_j) \approx \int log\{f(x)\}f(x)dx$
  2. Density approximation: $-\frac{1}{n}\sum_{j=1}^{n} log\hat{f}_n(X_j) \approx -\frac{1}{n}\sum_{j=1}^{n} logf(X_j)$
- Basic simulation studies showed integral approximation was acceptable for for moderately sized samples
- Will focus on improving the density approximation

Density Regression

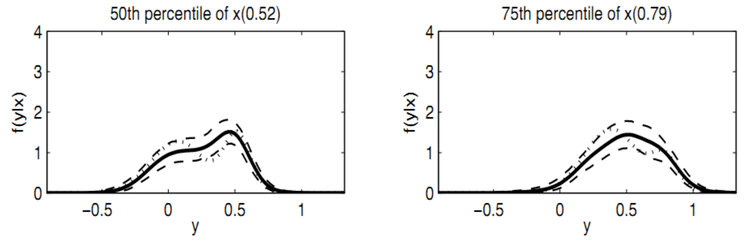
Technically, any density estimation method could be used to estimate S_i with. The two types worth considering are:

Nonparametric Density Estimation

- Includes both histogram estimators and kernel density estimators (KDE)
- Both methods can get arbitrarily close to any density - requires a lot of points, especially in multiple dimensions
- Estimation of joint density overlooks idea of inputs defining different output distributions - essentially marginalizes over input distribution

Bayesian Nonparametric Density Regression

- Density regression methods look for distributions that change with inputs - acknowledging relationship of *inputs values changing output density*
- Wide class of models, including *Kernel Stick Breaking Processes* (KSBP) and *Dirichlet Mixtures of Generalized Linear Models* [3]
- Methods are very flexible and can achieve almost density shape
- Not guaranteed to be more accurate with fewer points but fewer data causes high uncertainty posterior instead of just "poor estimates"
  - Might still be possible to conduct sensitivity analysis in this case!



Posterior predictive distribution from density regression model with KSBP prior [3]

Current and Future Work

- Currently trying to compute MII-type estimates from the out put of KSBP models
  - Need to determine asymptotic properties of these types of estimates - consistency, convergence, etc.
- Need further study on the information theoretic analogs for the interaction terms and the total sensitivity indices

References

- Beirlant, J., Györfi, E., van der Meulen, E.: Nonparametric entropy estimation: an overview (2001)
- Critchfield, G., Willard, K.: Probabilistic analysis of decision trees using monte carlo simulation. Medical Decision Making **6**(2), 85–92 (1986)
- Dunson, D., Park, J.: Kernel stick-breaking processes. Biometrika **95**(2), 307–323 (2008)
- Oakley, J., O'Hagan, A.: Probabilistic sensitivity analysis of complex models. Journal of the Royal Statistical Society: Series B **66**(3), 751–769 (2004)
- Sobol, I.: Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and computers in simulation **55**(1-3), 271–280 (2001)

Acknowledgments

This research is supported by NSF Science & Technology Center for Science of Information Grant CCF-0939370.

yjhuoh@stat.berkeley.edu