Sampling-Based Methods for Uncertainty and Sensitivity Analysis

Jon C. Helton
Sampling-Based Methods for Uncertainty and Sensitivity Analysis

- Involve generation and exploration of mapping from analysis inputs to analysis results

- Analysis input: $\mathbf{x} = [x_1, x_2, \ldots, x_n]$  

- Analysis results: $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_n(\mathbf{x})]$  

- Two Questions
  - What is the uncertainty in $\mathbf{y}(\mathbf{x})$ given the uncertainty in $\mathbf{x}$?
  - How important are the individual elements of $\mathbf{x}$ with respect to the uncertainty in $\mathbf{y}(\mathbf{x})$?
Uncertainty in Analysis Input

• Uncertainty in $y$ derives from uncertainty in $x$

• Assumption: Appropriate value for $y$ obtained if appropriate value for $x$ used

• Problem: Impossible to specify appropriate value for $x$ unambiguously

• Many possible values of $x$ of varying levels of plausibility

• Uncertainty with respect to $x$
  – Designated subjective or epistemic uncertainty
  – Characterized by distributions $D_1, D_2, \ldots, D_{nX}$ assigned to elements $x_1, x_2, \ldots, x_{nX}$ of $x$
Propagation of Uncertainty

• Generate sample: $x_k, k = 1,2,\ldots,nS$

• Evaluate $y$: $y(x_k), k = 1,2,\ldots,nS$

• Resultant mapping: $[x_k, y_k(x_k)], k = 1,2,\ldots,nS$

• Mapping forms basis for
  – Uncertainty analysis (distribution functions, box plots…)
  – Sensitivity analysis (scatterplots, regression analysis,…)
Components of Sampling-Based Uncertainty/Sensitivity Analysis

- Characterization of uncertainty in \( x \) (i.e., definition of \( D_1, D_2, \ldots, D_{nX} \))

- Generation of sample from \( x \) (i.e., generation of \( x_k, \ k = 1,2,\ldots,nS \), in consistency with \( D_1, D_2, \ldots, D_{nX} \))

- Propagation of sample through analysis (i.e., generation of mapping \( [x_k, y(x_k)], \ k = 1,2,\ldots,nS \))

- Presentation of uncertainty analysis results (i.e., approximations to the distributions of the elements of \( y \) obtained from \( y(x_k), \ k = 1,2,\ldots,nS \))

- Determination of sensitivity analysis results (i.e., exploration of the mapping \( [x_k, y(x_k)], \ k = 1,2,\ldots,nS \))
Example Analysis: Two-Phase Fluid Flow

• Context: Radioactive waste disposal facility in bedded salt
• Mathematics: System of nonlinear partial differential equations
• Numerics: Finite difference procedure on two-dimensional grid
• Uncertainties: 31 uncertain inputs (i.e., $nX = 31$)
Example Analysis: Uncertain Inputs

• Uncertain inputs
  – $ANRGSSAT$-Residual gas saturation in anhydrite
  – $BHPRM$-Logarithm of borehole permeability
  – $BPCOMP$-Logarithm of bulk compressibility of brine pocket
  Correlation: $-0.75$ rank correlation with $BPPRM$
  ...
  – $WRBRNSAT$-Residual brine saturation in waste
  – $WRGSSAT$-Residual gas saturation in waste

• Vector representation
  $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}], \ nX = 31$
  $= [ANRGSSAT, BHPRM, \ldots, WRGSSAT]$
Example Analysis: Results

- Many analysis results
  - Pressure
  - Brine and gas saturations
  - Gas production due to corrosion and microbial degradation of cellulose
  - Brine and gas flows across specified boundaries
  - and many more

- \( \mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_n(\mathbf{x})] \)
  - \( nY \) potentially very large
  - Elements \( y_j(\mathbf{x}) \) of \( \mathbf{y}(\mathbf{x}) \) functions of time and space, i.e.,
    \( y_j(x,y,t,\mathbf{x}) \)
Characterization of Uncertainty in $\mathbf{x}$

- Corresponds to definition of $D_1, D_2, \ldots, D_{nX}$

- Most important part of sampling-based uncertainty/sensitivity analysis

- Determines both
  - Uncertainty in $\mathbf{y}$
  - Sensitivity of $\mathbf{y}$ to elements of $\mathbf{x}$

- $D_1, D_2, \ldots, D_{nX}$ typically defined through expert review/elicitation process

- Extensive literature available

- Not primary focus of this presentation
Observation on the Characterization of Uncertainty in \( \mathbf{x} \)

- Understand type of information characterized by \( D_1, D_2, \ldots, D_{nX} \)
  - Degree of belief with respect to where appropriate value of each element of \( \mathbf{x} \) is located for use in analysis

- Do not confuse uncertainty with spatial, temporal or experimental variability
  - If analysis uses a spatially or temporally averaged value for \( x_j \), then \( D_j \) should characterize uncertainty in this average, \textbf{not} the variability that is averaged over
  - Similarly, experimental variability is not the same as uncertainty in an analysis input derived from variable experimental outcomes
Possible Construction Procedure for $D_j$

- Specify
  - Min and max values
  - Median
  - Quartiles
  - Additional quantiles

- Document rationale
Possible Construction Procedure for $D_j$ with Multiple Experts
Observations on Construction of $D_1, D_2, \ldots, D_{nX}$

- Can be most expensive and important part of analysis

- Care and effort dependent on
  - Purpose of analysis
  - Time and resources available

- Exploratory analysis
  - Crude specifications (e.g., uniform and loguniform)
  - Will probably reveal dominant variables

- Analysis that forms basis for important decisions
  - Care required in construction of $D_1, D_2, \ldots, D_{nX}$
  - $D_1, D_2, \ldots, D_{nX}$ influence both uncertainty and sensitivity analysis results
Possible Analysis Strategy

• Perform initial (exploratory) analysis
  – Crude definitions for $D_1, D_2, \ldots, D_{nX}$
  – Identify important variables with sensitivity analysis

• Characterize uncertainty in important variables

• Perform second analysis
  – Presentation or decision-aiding analysis
  – Use refined distributions for important variables
Example Analysis: Distributions

- Distributions developed from 31 uncertain variables
- Examples below
Sampling Procedures: Random Sampling

- Random sample: $\mathbf{x}_k = [x_{1k}, x_{2k}, \ldots, x_{nX,k}], \quad k = 1, 2, \ldots, nR$

- Sample elements (i.e., $\mathbf{x}_k$’s) from different regions of sample space occur in direct relationship to the probability of these regions

- Each sample element selected independently of all other sample elements
Sampling Procedures: Importance Sampling

• Importance sample: \( \mathbf{x}_k = [x_{1k}, x_{2k}, \ldots, x_{nX,k}] \), \( k = 1,2,\ldots,nS \)

• Sample space divided into strata \( S_1, S_2, \ldots, S_{nS} \) which
  – Typically have unequal probabilities
  – Assure inclusion of specific regions of sample space in analysis

• Sample element \( \mathbf{x}_k \) randomly sampled from strata \( S_k \)
Sampling Procedures: Latin Hypercube Sampling

- Latin hypercube sample (LHS): $x_k = [x_{1k}, x_{2k}, \ldots, x_{nk}, k = 1, 2, \ldots, nLHS$

- Generation of sample
  - Range of each $x$ divided in $nLHS$ intervals of equal probability
  - Value for $x_j$ (i.e., $x_{jk}$) randomly selected from each interval
  - Values for $x_1$ randomly paired without replacement with values for $x_2$ to produce $nLHS$ pairs
  - Preceding pairs randomly combined without replacement with values for $x_3$ to produce $nLHS$ triples
  - Process continues through all variables to produce $nLHS$ sample elements
Latin Hypercube Sampling: Example

$RU_1$: $U_1 = -0.37$
$RU_2$: $U_2 = -0.20$

$RU_5$: $U_5 = 0.49$

$RV_1$: $V_1 = 0.74$
$RV_2$: $V_2 = 1.07$

$RV_3$: $V_5 = 3.39$
Latin Hypercube Sampling: Example (cont)

\[ x_1 = [U_1, V_5] \]
\[ x_2 = [U_2, V_1] \]
\[ \ldots \]
\[ x_5 = [U_5, V_4] \]

\[ x_1 = [U_1, V_3] \]
\[ x_2 = [U_2, V_2] \]
\[ \ldots \]
\[ x_5 = [U_5, V_1] \]
Comparison of Sampling Techniques

• Random sampling preferred when sufficiently large samples are possible
  – Easy to implement
  – Easy to explain
  – Unbiased estimates for means, variances and distribution functions
  – Sufficiently large samples may not be possible

• Importance sampling
  – Used when random sampling not computationally feasible for estimation of extreme quantiles
  – Development of strata and strata probabilities often challenging
  – Requires a priori knowledge about problem
  – Event trees are algorithms for defining importance sampling
Comparison of Sampling Techniques (cont)

• Latin hypercube sampling
  – Unbiased estimates for means and distribution functions
  – Dense stratification across range of each variable
  – Used when large samples not computationally practicable and estimation of high quantiles not required
  – Preceding is typically the case in uncertainty/sensitivity analyses to assess effects of subjective uncertainty
  – Uncertainty/sensitivity results robust with relatively small sample sizes (e.g., $n_{LHS} = 50$ to $200$)
Random and LH Sampling: Correlation Control

- Correlation control important
  - Correlated variables should have appropriate correlations
  - Uncorrelated variables should have correlations close to zero

- Imposition of complex correlation structure not easy

- Iman and Conover have developed method to impose rank correlations
  - Distribution free
  - Works with random and LH sampling
  - Preserves intervals used in LH sampling
Correlation Control: Example

- Left graph: Rank Correlation of 0.00
  - $X_1$: Normal with Mean = 0 and Var = 1
  - $X_2$: Uniform on (0,1)

- Right graph: Rank Correlation of 0.25
  - $X_1$: Normal with Mean = 0 and Var = 1
  - $X_2$: Uniform on (0,1)
Correlation Control: Example (cont)

- Left diagram: Rank CORRELATION OF 0.50
  - $X_1$: NORMAL WITH MEAN = 0 AND VAR = 1
  - $X_2$: UNIFORM ON (0,1)

- Right diagram: Rank CORRELATION OF 0.75
  - $X_1$: NORMAL WITH MEAN = 0 AND VAR = 1
  - $X_2$: UNIFORM ON (0,1)
Correlation Control: Example (cont)
Example Analysis: Generation of Sample

• LH sampling used

• $nLHS = 300$ samples from $nX = 31$ variables

• Sample: $\mathbf{x}_k = [x_{1k}, x_{2k}, \ldots, x_{31k}]$, $k = 1, 2, \ldots, 300$

• Tests for robustness
  – Sample generated as 3 independent replicated samples of size 100 each
  – Individual replicated samples

    \begin{align*}
    \text{R1: } & \mathbf{x}_k, \quad k = 1, 2, \ldots, 100 \\
    \text{R2: } & \mathbf{x}_k, \quad k = 1, 2, \ldots, 100 \\
    \text{R3: } & \mathbf{x}_k, \quad k = 1, 2, \ldots, 100
    \end{align*}

• Iman/Conover technique used to control correlations
Propagation of Sample through Analysis

• Generates mapping \([x_k, y(x_k)], k = 1, 2, \ldots, n_S\)

• Requires evaluation of \(y(x_j)\)
  – Details analysis specific
  – Can be most computationally demanding part of analysis

• Example analysis
  – Required 4-5 hrs CPU time per model evaluation on VAX Alpha
  – Produced large quantity of temporally and spatially variable results
Uncertainty Analysis: Scalar Results

• Single scalar result: \( y_k = y(x_k), \ k = 1, 2, \ldots, nS \)

• Mean and variance

\[
\hat{E}(y) = \frac{\sum_{k=1}^{nS} y_k}{nS}, \quad \hat{V}(y) = \frac{\sum_{k=1}^{nS} \left[ y_k - \hat{E}(y) \right]^2}{(nS - 1)}
\]
Uncertainty Analysis: Scalar Results (cont)

- Distribution function
Example Analysis: CDFs and CCDFs

CDF: Probability Value ≤ y

CCDF: Probability Value > y

y = E0: WAS_PRES, Pa (10^6)

TRI-6942-6041-0
Example Analysis: CDF and Density Function
Example Analysis: Box Plots

BRAGFLO (E0, R1, R2, R3)

MB 138 (N) BRM38NIC
MB 138 (S) BRM38SIC
Anh A, B (N) BRAABNIC
Anh A, B (S) BRAABSIC
MB 139 (N) BRM39NIC
MB 139 (S) BRM39SIC
MB Tot BRAALIC
Rep Tot BRNREPTC

Brine Flow (m^3)

Time: 10000 yr

Min {1.5x Box, Largest Obs}

Key: 25th Percentile Median Mean 75th Percentile Extreme Obs

TRI-6342-4830-0
Uncertainty Analysis: Functions

- Analysis outcomes are often functions of one or more variables
- Uncertain analysis inputs result in many possible values for such functions
- Example analysis: pressure as a function of time
Example Analysis: Pressure at Time $t$

![Graphs showing pressure over time for two scenarios: BRAGFLO (E0, R1) and BRAGFLO (E0, R1, R2, R3). The graphs depict the pressure changes with time for the lower panel of the WAS_PRES dataset.](image)
Example Analysis: Replicated Samples and Stability
Sensitivity Analysis

- Involves exploration of mapping \([x_k, y(x_k)], k = 1,2,\ldots, nS\)
- Available techniques for sensitivity analysis
  - Examination of scatterplots and cobweb plots
  - Correlation and partial correlation analysis
  - Regression analysis
  - Stepwise regression analysis
  - Rank transforms to linearize monotonic relationships
  - Nonparametric regression: Loess, additive models, projection pursuit, recursive partitioning
  - Tests for patterns based on gridding: nonmonotonic relations, nonlinear relations
Sensitivity Analysis (Continued)

- Tests for patterns based on distance measures
- Multidimensional Kolmogorov-Smirnov test
- Tree-based searches
- Squared differences of ranks
- Top-down concordance with replicated samples
- Variance decomposition
Examination of Scatterplots

- Simplest sensitivity analysis technique
- Points \((x_{jk}, y_k), k = 1,2,\ldots,nS,\) plotted for each \(x_j\) in \(\mathbf{x}\)
- Resulting plots visually examined for patterns
- Effective with LH sampling due to full stratification over range of each \(x_j\)
- Examples follow
Example Analysis: Scatterplots

BRAGFLO (E2 at 1000 yr, R1, R2, R3)

Borehole Permeability (m²) $10^x$, x = BHPRM

Time: 10000 yr
Stepwise Regression: Ranking Variable Importance

- Order of selection in stepwise procedure
- Changes in $R^2$ values at successive steps
- Absolute values of SRCs
- With independent sampled variables, three preceding ranking procedures produce same ordering of variable importance
Example Analysis: Pressure at Time $t$
Example Analysis: Compact Representation of Stepwise Regression

- Compact Summary of Stepwise Regression Analyses for Pressure in the Repository at 10,000 yr Under Undisturbed Conditions (i.e., \( y = E0:WAS\_PRES \) at 10,000 yr)

<table>
<thead>
<tr>
<th>Step (^a)</th>
<th>Variable (^b)</th>
<th>SRC (^c)</th>
<th>( R^2 ) (^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WMICDFLG</td>
<td>0.718</td>
<td>0.508</td>
</tr>
<tr>
<td>2</td>
<td>HALPOR</td>
<td>0.466</td>
<td>0.732</td>
</tr>
<tr>
<td>3</td>
<td>WGRCOR</td>
<td>0.246</td>
<td>0.792</td>
</tr>
<tr>
<td>4</td>
<td>ANHPRM</td>
<td>0.129</td>
<td>0.809</td>
</tr>
<tr>
<td>5</td>
<td>SHRGSAT</td>
<td>0.070</td>
<td>0.814</td>
</tr>
<tr>
<td>6</td>
<td>SALPRES</td>
<td>0.063</td>
<td>0.818</td>
</tr>
</tbody>
</table>

\(^a\) Steps in stepwise analysis.

\(^b\) Variables listed in the order of selection in regression analysis.

\(^c\) Standardized regression coefficients (SRCs) for variables in final regression model.

\(^d\) Cumulative \( R^2 \) value with entry of each variable into regression model.
Example Analysis: Time-Dependent SRCs and PCCs
Identification of Nonmonotonic Patterns

- Regression-based sensitivity analyses sometimes perform poorly

- Rank transformations possible alternative
  - Linearizes monotonic relationships
  - Will not help for nonmonotonic relationships

- Example follows
Example Analysis: Failure of Regression-Based Techniques
Example Analysis: Failure of Regression-Based Techniques (cont)

• Examination of scatterplots identified dominant variable
Test for Nonmonotonic Patterns

• Three tests
  – $F$-test for common means (CMNs)
  – $\chi^2$-test for common medians (CMDs)
  – Kruskall-Wallis test for common locations (CLs)

• Tests based on
  – Subdividing values of $x_j$ into intervals
  – Testing to determine if $y$ has common measure of central tendency across these intervals
Tests for Nonmonotonic Patterns: Subdivision of $x_j$
Identification of Random Patterns

• Determine if points in scatterplot appear to be random conditional on distributions for $x_j$ and $y$

• Based on dividing points into two-dimensional grid

• $\chi^2$-test used to indicate if scatterplot appears to be random

• Denoted test for statistical independence (SI)
Identification of Random Patterns: Subdivision of $x_j, y$
Additional Information

