Input Uncertainty and Potential-to-Validate: Sampling Plans for Monte Carlo Assessment

Max D. Morris
Department of Statistics
Department of Industrial and Manufacturing Systems Engineering
Iowa State University

Lisa Moore, Mike McKay
Statistical Sciences Group
Los Alamos National Laboratory

Knoxville, TN, June 7-9, 2006

Context

- Deterministic Model: $y = \mathcal{M}(\mathbf{x})$
  - $\mathbf{x}$: *input*, k-dimensional where k is sometimes large, defining a specific setting/problem/situation ...
  - y: *output*, the result
- Reality: an observed y^R resulting from setting $\mathbf{x}^R$
- Idealized Validation of $\mathcal{M}$:
  - Would y^R match $\mathcal{M}(\mathbf{x}^R)$, if we had y^R and knew $\mathbf{x}^R$?
  - Or, would

$$\Delta(y^R, \mathcal{M}(\mathbf{x}^R))$$

be "small enough", for an appropriate non-negative loss Δ?

With Measurements, but Uncertain Setting

- If we had y^R but didn't know $\mathbf{x}^R$ ...
- Express uncertainty about $\mathbf{x}^R$ by treating $\mathbf{x}$ as a random variable with a specified distribution.
- Would

$$E_{\mathbf{x}}\{\Delta(y^R, \mathcal{M}(\mathbf{x}))\}$$

be "small enough"?

With Neither Certainty nor Measurements

- Evaluate *Potential-to-Validate* with the most optimistic "fake data" for a given setting.
- Would

$$\underline{\underline{E\Delta}} \;=\; min_{y^R}\; E_{\mathbf{x}}\{\Delta(y^R, \mathcal{M}(\mathbf{x}))\}$$

  be "small enough"?
- e.g., for squared-error loss:

$$\underline{\underline{E\Delta}} \;=\; \theta\, Var_{\mathbf{x}}\{\mathcal{M}(\mathbf{x})\}$$

  the focus of variance-based Uncertainty Analysis (e.g. Saltelli), but the idea works for other Δ.

Where to Spend Effort Before Validation

- Let $\mathbf{x} = (x_i, \mathbf{x}_{(i)})$. What *would* the potential-to-validate be if uncertainty about x_i were eliminated?

$$min_{y_i^R}\ E_{\mathbf{x}_{(i)}|x_i}\{\Delta(y_i^R, \mathcal{M}(x_i, \mathbf{x}_{(i)}))\}$$

- Or, in the current absence of certainty about x_i,

$$\underline{\underline{E\Delta}}_{(i)} \ = \ E_{x_i}\ min_{y_i^R}\ E_{\mathbf{x}_{(i)}|x_i}\{\Delta(y_i^R, \mathcal{M}(x_i, \mathbf{x}_{(i)}))\}$$

- e.g., for squared-error loss:

$$\underline{\underline{E\Delta}}_{(i)} \ = \ \theta\ E_{x_i}\ Var_{\mathbf{x}_{(i)}|x_i}\{\mathcal{M}(x_i, \mathbf{x}_{(i)})\} \ = \ \theta\ T_{(i)}$$

where $T_{(i)}$ is expected conditional variance, or "total variance" (e.g. Saltelli) associated with $\mathbf{x}_{(i)}$.

Simulation-Based Estimation of $\underline{\underline{E\Delta}}$ and $\underline{\underline{E\Delta}}_{(i)}$

- Will need independent distn's for each input.
- $\underline{\underline{E\Delta}}_{(i)}$ = average $\underline{\underline{E\Delta}}$ with x_i fixed and $\mathbf{x}_{(i)}$ varying, where "average" is over randomly chosen values of x_i.
- Will need groups of model runs that have the same (randomly chosen) x_i value, but different (randomly chosen) values for the other inputs.
- Potential-to-Validate is most improved by eliminating uncertainty in inputs x_i for which $\underline{\underline{E\Delta}}_{(i)}$ is smallest (equivalently, $T_{(i)}$ for squared-error loss).

"Substituted Column Arrays"

Sobol': 1990, 1993 translation

$$\begin{bmatrix} \cdot & \cdot & \cdot \\ \mathbf{x}_1^1 & \mathbf{x}_2^1 \; \ldots & \mathbf{x}_k^1 \\ \cdot & \cdot & \cdot \end{bmatrix}_n \begin{matrix} \leftrightarrow \\ \leftrightarrow \\ \leftrightarrow \end{matrix} \begin{bmatrix} \cdot & \cdot & \cdot \\ \mathbf{x}_1^1 & \mathbf{x}_2^2 \; \ldots & \mathbf{x}_k^2 \\ \cdot & \cdot & \cdot \end{bmatrix}_n \begin{matrix} \leftrightarrow \\ \leftrightarrow \ldots \\ \leftrightarrow \end{matrix} \begin{bmatrix} \cdot & \cdot & \cdot \\ \mathbf{x}_1^1 & \mathbf{x}_2^a \; \ldots & \mathbf{x}_k^a \\ \cdot & \cdot & \cdot \end{bmatrix}_n \begin{matrix} \rightarrow \widehat{\underline{\underline{E\Delta}}}|x_1 \\ \rightarrow \widehat{\underline{\underline{E\Delta}}}|x_1 \\ \rightarrow \widehat{\underline{\underline{E\Delta}}}|x_1 \\ \downarrow \\ \widehat{\underline{\underline{E\Delta}}}_{(1)} \end{matrix}$$

- d.f. $= n \times (a-1) \qquad N = n \times (1 + k \times (a-1))$
- "efficiency" $\approx k^{-1}$, d.f. per model evaluation
- $\underline{\underline{E\Delta}}$ is estimable from the runs in any one array

"Substituted Column Arrays"

Example: k=3 inputs, a=2 arrays, n=4 runs/array, $x_i \sim$ unif[0,1]

$\Delta = (y^R - \mathcal{M}(\mathbf{x}))^2$, $\underline{E\Delta}_{(i)}$ = average $S^2(y)$ ...

x_1	x_2	x_3	y		x_1	x_2	x_3	y		$S^2(y)$
.03	.11	.82	0.96	$\leftrightarrow$	**.03**	.53	.09	0.65	$\rightarrow$	.0240
.11	.06	.69	0.86	$\leftrightarrow$	**.11**	.76	.13	1.00	$\rightarrow$	.0049
.37	.87	.46	1.70	$\leftrightarrow$	**.37**	.77	.84	1.96	$\rightarrow$	.0169
.79	.51	.58	1.88	$\leftrightarrow$	**.79**	.41	.37	1.57	$\rightarrow$	.0240
										$\downarrow$
										.0175

- $\hat{T}_{(1)} = .0175$

"Balanced Replication Arrays"

Based on the pattern of a Balanced Incomplete Block Design, where each pair of treatments appear together in exactly one block, e.g.

$$
\begin{array}{ccc}
 & \text{“blocks”} & \\
\text{“treatments”} & \left[\begin{array}{ccccccc}
\bullet & & & & \bullet & & \bullet \\
\bullet & \bullet & & & & \bullet & \\
 & \bullet & \bullet & & & & \bullet \\
\bullet & & \bullet & \bullet & & & \\
 & \bullet & & \bullet & \bullet & & \\
 & & \bullet & & \bullet & \bullet & \\
 & & & \bullet & & \bullet & \bullet
\end{array}\right]_{n=7} & \begin{array}{l}
\leftarrow \\
\leftarrow \quad \widehat{\underline{E\Delta}}|x_1 \\
 \\
\leftarrow \\
 \\
 \\

\end{array}
\end{array}
$$

"Balanced Replication Arrays"

- construct a independent arrays in this pattern, average $\widehat{\underline{E\Delta}}|x_i$ values, one from each array.
- d.f. $= a \times (u-1)$, where u is the blocksize of the BIBD
- $N = a \times n$
- "efficiency" $\approx k^{-\frac{1}{2}}$
- $\underline{E\Delta}$ is estimable from the corresponding runs across arrays

Example: Species Competition

- System of 8 equations for 8 species densities:

$$N_1'(t) = N_1(t)\tfrac{r_1}{K_1}[\ (K_1 - N_1(t)) - \textstyle\sum_{m\neq 1} \alpha_{1m} N_m(t)\ ]$$

$$N_i'(t) = N_i(t)\tfrac{r_i}{K_i}[\ (K_i - N_i(t)) - \textstyle\sum_{m\neq i} \alpha_{im} N_m(t)\ ] - \beta N_i(t), \quad i \neq 1$$

$$\alpha_{i,j} = 1.2,\ j > i+1 \qquad \alpha_{i,j} = 0.2,\ j \leq i+1$$

$$r_i = 1.0,\ i = 1...8 \qquad \beta = 0.5$$

species(i)	1	2	3	4	5	6	7	8
K_i	20	7	6	5	4	3	2	1

Example: Species Competition

- $N_1(0) \;=\; 0.5\;, \quad x_i = N_{i+1}(0) \sim \text{unif}[0.1, 0.2],\; i = 1...7$
- $y \;=\; N_1(5)$
- Asymmetric Linear Loss:

$$\Delta(y^R, \mathcal{M}(\mathbf{x})) = \begin{array}{rl} [\mathcal{M}(\mathbf{x}) - y^R], & \mathcal{M}(\mathbf{x}) > y^R \\ 2 \times [y^R - \mathcal{M}(\mathbf{x})], & \mathcal{M}(\mathbf{x}) \leq y^R \end{array}$$

- For $y_j \;=\; \mathcal{M}(x_i, \mathbf{x}_{(i),j}) \quad j = 1 \ldots m$

$$argmin_{y^R}\; ave_{y_j}\; \Delta(y^R, y_j) = y^* \quad \in \quad \{y_1, y_2, y_3, ..., y_m\}$$

- Estimated $\underline{\underline{E\Delta}}|x_i$:

$$\frac{1}{m}[\sum_{y_j > y^*} (y_j - y^*) + 2 \sum_{y_j < y^*} (y^* - y_j)]$$

Example: Species Competition

Example: k=7 inputs, 7 runs/array, one of a=50 arrays

x_1	x_2	x_3	x_4	x_5	x_6	x_7	y	
.18397	.11297	.14310	.15244	**.15358**	.11940	**.13649**	13.882	←
.18397	**.13505**	.17935	.18486	.14569	**.10225**	.19577	13.696	←
.18623	**.13505**	**.11565**	.10929	.14873	.16611	**.13649**	13.954	
.18397	.14136	**.11565**	**.10374**	.11559	.14043	.10785	14.006	←
.18330	**.13505**	.14591	**.10374**	**.15358**	.19171	.18767	13.872	
.13476	.15396	**.11565**	.15036	**.15358**	**.10225**	.18158	13.864	
.11169	.11994	.18742	**.10374**	.17219	**.10225**	**.13649**	13.856	

- must repeat this pattern, with new random draws for a arrays
- $\underline{\widehat{E\Delta}}_{(1)}$ computed as the average of $\underline{\widehat{E\Delta}}|x_1$ values from groups as depicted above, et cetera.
- standard errors are the corresponding standard deviations of $\underline{\widehat{E\Delta}}|x_1$ values, divided by $\sqrt{a}$.

Example: Species Competition

- $\underline{\widehat{E\Delta}} = 0.2592$, an average of 7 indices each based on m=50.
- $\underline{\widehat{E\Delta}}_{(i)}$, each an average of 50 indices each based on m=3.

species	2	3	4	5	6	7	8
$\underline{\widehat{E\Delta}}_{(i)}$	0.1005	**0.0746**	0.0898	0.0875	0.1109	0.1047	0.0938
std. err.	0.0076	0.0063	0.0066	0.0066	0.0081	0.0078	0.0072

- units = N

Summary and Conclusion

- Potential-to-Validate, in the presence of input uncertainty, is directly related to probabilistic S/U analysis.
- BIBD-based input sampling arrays are an alternative to SCA's that are more efficient in many cases:
  - Morris, M., L. Moore, and M. McKay. Sampling Plans Based on Incomplete Block Designs for Evaluating the Importance of Computer Model Inputs. *Journal of Statistical Planning and Inference*, in press.
- Overview of sampling-based methods in S/U analysis:
  - Saltelli, A., K. Chan, and M. Scott (eds.) *Sensitivity Analysis*. John Wiley and Sons, New York.