NSF Simulation Workshop Special Interests Group Discussion: Modeling Error

July 25, 2012

1 Group members

Nan Chen (inscriber), Peter Frazier (leader), Peter Glynn, Jeff Hong, Hai Lan, Guangwu Liu, Szu Hui Ng, Xiaoqun Wang, Xiaowei Zhang

2 Monday Discussion

The group had general discussions on the sources of modeling error, potential solutions, and corresponding applications. Following sources of uncertainty are considered as important:

- input uncertainty
- initial condition uncertainty
- model uncertainty
  - simulation dynamics
  - input distribution
- objective function uncertainty

These uncertainties might be studied depending on what data are available. Several research topics have also been identified along this line.

Input data	Output data	Remarks
No	No	
Yes	No	
No	Yes	censored data, filtering methods
Yes	Yes	executing systems

2.1 Filtering problem

Filtering methods are used to estimate the hidden state from related observations, which might be useful in the scenarios where input data or initial conditions are not available. Mathematically, (Z_t, X_t) is a stable process described or generated by the discrete event systems. X_t is unobservable, but Z_t is observable. Filtering is to compute the conditional distribution $[X_t|Z_{1:t}]$ of the hidden state at time t given historical observations of Z_τ up to time t.

Challenges: How to effectively implement filtering (with supercomputers)? Typically the conditional distribution has very high dimension, and the numerical computation is difficult to solve. Not much attention are paid from researchers in the stochastic models.

2.2 Model uncertainties

Several related topics have been proposed to quantify or handle model uncertainties in simulation.

1. Model risk. If a set of candidates models are possible, how to select the "correct" one from the set, and how to measure the model risks of picking the wrong models. *Bayesian* approach is mentioned as one way to mitigate the risk.

2. Robust simulation. In the presence of model uncertainty, how to run the simulation such that the simulation can still represent the true response surface robustly? How can be use/modify robust optimization techniques in simulation to achieve this goal?

Both robust formulation (worst case scenario, etc) and Bayesian decision theoretic approaches were mentioned as possible ways to handle model uncertainties.

2.3 Input uncertainty

Three issues were discussed on the input uncertainty.

1. Simulation with uncertain input distribution. Jeff mentioned his work on robust simulation using change-of-measure method, where uncertainty in input distribution can be taken care of.

2. Robust estimation. Given the input data, how to robustly estimate the input distribution, especially the extreme tails of the distribution?

3. Nonparametric sampling. If the input data were used non-parametrically, how to sample from them sensibly such that different scenarios (e.g., different mean, variance) can be obtained?

2.4 Big data

When a lot of data are available, they will reject every hypothesis. What can we do to utilize these data? Another issue is how to handle the big data given the constraints of the computation speed and storage space. How to decompose and parallel the operations on the big data is of interests.

2.5 Being humble

The famous statistician George Box once said "All models are wrong, some are useful". We should be humble about the results obtained from the simulation. For example, how do we know whether a rare event probability 10^-7 is meaningful or not? Since behavior may change in tails, assuming Pareto or exponential or other tails in the areas without data might be dangerous. Careful interpretation or uncertainty quantification might be needed to avoid misleading conclusions (e.g., the role of copula in financial crisis).

3 Tuesday's Discussion

We further discussed the issues in model uncertainty and parameter uncertainty in Tuesday's session.

3.1 Model uncertainty

The issue discussed is whether the model bias will influence the optimal solution in the simulation optimization problems. In the special case of SA algorithms, if the function is twice continuously differentiable, the bias can be expressed in a analytical way. In general, the model bias

- might not pose challenges if only the optima not the optimal value is of interests
- bias might be estimated due to model uncertainty, Jack-knife estimator

Another topics is when knowledge about the system is available, how to utilize this information with the data available. Put a prior on the mean function, and a Gaussian process prior on the model bias. The Bayesian approach may handle this problem well.

3.2 Parameter uncertainty

The estimation of the input parameters should depend on how these parameters will be used in the later stages. Most of the statistical estimators focus on the central part of the distribtuion, which may lead to diaster in the tails. Some examples include

- MLE estimator of the traffic intensity in M/M/1 queues may lead to large discrepancy in queue length distributions.
- If the parameters in the linear programming are estimated, the uncertainty in the optimal solution might be exemplified.

Jeff mentioned the idea of using simulation as a learning tool to calibrate the parameters. Essentially the performance of the system output is matched rather than the input parameters per se. This calibration is related to the decision theoretical formulation of statistical estimation. In other words, we can define the loss function of the estimator by taking consideration of how it gonna be used: $L(\hat{\theta}) = ||g(\hat{\theta}) - g(\theta)||^2$, where $g(\cdot)$ is the function linking the parameter with the performance measure of interests. However, this decision theoretic formulation might be difficult to solve even when $g(\cdot)$ is known. Potential applications include

- Estimation of the input parameters in the queueing system

- Estimation of volatility using stock prices. Paper by **Macro Avellaneda**

4 Topics to present

Following are some potential topics to be presented to the workshop attendees.

1. Type of uncertainty
2. Unifying models to account for different uncertainties

   $$g(x) = f(x, \theta, \beta, \omega) + h(x, \cdot) + \epsilon(x)$$

   where $g(x)$ is the true response, x is the design variables, θ is partially observed parameters, β is unobservable parameters, ω is the randomness in the simulation model, $h(\cdot)$ is the model bias, and $\epsilon(x)$ is the observation error or inherent randomness. In a simple case, we want to find the parameters

   $$\hat{\theta}, \hat{\beta} = \arg\min_{\theta,\beta} \mathrm{E}[f(x, \theta, \beta, \omega)] - g(x)$$

3. .[*simulation*] Filtering: estimate the hidden state
4. .[*simulation*] parameter uncertainty: combine estimation with calibration; integrate the use of the parameters to adjust the parameter estimation
5. .[*generic*] uncertainty quantification
6. .[*generic*] how to combined data with physical knowledge of the simulation model.
7. .[*generic*] Extrapolation vs interpolation: depending on the distance from the existing data, and is related to multi-fidelity data.
8. trace data: computational questions–cannot use it all. How to determine what data is most representative.

9. How to analyze the simulation when the input is composed of both trace data and synthetic input? Questions include how to construct the confidence interval of the simulation output given limited data.

10. linking to optimization

    - optimizer curse
    - objective uncertainty
    - related to loss function
    - uncertainty aware optimization, robust optimization

Modeling Error: Special Interest Group 2

Nan Chen (inscriber), Peter Frazier (leader), Peter Glynn, Jeff Hong, Hai Lan, Guangwu Liu, Szu Hui Ng, Xiaoqun Wang, Xiaowei Zhang

NSF Simulation Workshop Special Interest Group 2

Wednesday July 25, 2012
Shanghai
NSF Simulation Workshop Special Interest Group 2

Outline

1. Types of Modeling Error

2. Conversations and Research Questions

Types of Modeling Error

- Parameter Uncertainty
- Model Uncertainty
  - Input Distribution Uncertainty
  - System Dynamics Uncertainty
- Initial Condition Uncertainty
- Objective Function Uncertainty

Parameter Uncertainty

Our simulation model has some parameters, whose values we don't know.

- Example: We don't know the arrival rate in a queueing model.
- Example: We don't know the rate of infection in an epidemic model.
- Example: We don't know the volatility in a financial model.
- There can be data on the input side, the output side, or both.

Model Uncertainty: Input Distribution

We do not know the probability distribution of the inputs to our simulatino model.

- Example: We don't know the distribution governing the arrival process.
- Example: We don't know the distribution of demand in a supply chain model.
- There is some overlap with parameter uncertainty, since a non-parametric model for the input distribution is really just a parametric model with an unbounded number of parameters.

Model Uncertainty: System Dynamics

Our simulation does not include full the dynamics of the real world, because of simplifications made in designing the model.

- Example: customers may decide whether or not to abandon a queue based on the number of people in the queue, but our model assumes they have an exogeneously generated patience time.
- George Box: All models are wrong, some are useful.

Initial Condition Uncertainty

We cannot measure the full state of the real system at the time at which we would like simulated time to begin.

- Example: Weather and Climate models.
- Example: Agent-based economic models.
- Example: Epidemic models.
- This is particularly a problem if the system exhibits chaotic behavior (strong sensitivity to initial conditions).

Objective Uncertainty

We have multiple objectives in a simulation optimization problem, and we are not sure what the decision-maker's preferences over these objectives is. More generally, the objective we care about can be some partially unknown transformation of something we can simulate.

- Example: Cost vs. quality of care in a healthcare simulation.
- Example: Design of cardiovascular bypass grafts, where we can simulate stress on the wall of the artery, and we care about patient mortality.
- This type of uncertainty is particular to simulation optimization.
- Strategy: produce a Pareto frontier and show it to the decision-maker.

Outline

1. Types of Modeling Error
2. Conversations and Research Questions

Consider your loss function when estimating parameters

- A common strategy for parameter uncertainty, when there is input data, is to use a standard point estimator, e.g., the MLE, or linear regression.
- This ignores the way that we will use the estimate.
- In a perfect world, we would use a statistical technique based on our actual loss function.
- Example: in a newsvendor problem with covariate dependent demand distribution, our loss function is the amount of money+goodwill we will lose by misestimating tomorrow's demand, and in a perfect world we would use a statistical technique that knows this.
- Example: if our simulation model depends most strongly on the right tail of an input distribution, we should use an estimation technique that tries hard to get the tails right.

Consider your loss function when estimating parameters

- Challenge: In the real world, our loss function depends in a complicated way on the simulation dynamics, and the way that we will use the results from the simulation, and is hard to include.
- Research question: How can we incorporate our loss function in a tractable way?
- Advice: think about what your statistical technique is optimizing, and what you are optimizing, and whether they are compatible.

How can we combine data obtained on the input and output side?

- Consider a problem with parameter uncertainty, and no other kinds of uncertainty.
- For example, a queueing simulation of a hospital emergency ward.
- If we only have observations on the input data (e.g., a record of arrivals), we fit a parametric model of this input distribution, with some residual uncertainty.
- If we only have observations on the output data, we solve a calibration problem in which we change the parameters to fit the output of the simulation to the data.
- Research question: If we have both input data and output data, how can we use both datasets to estimate the parameters?

How can we combine data obtained on the input and output side?

Initial thoughts:

- In the Bayesian setting, it is conceptually clear that we should find the MAP estimator, but this is not possible exactly because we cannot get the likelihood of the observed data (as a function of the input parameters) in closed form.
- It should also be possible to study this question with system dynamics model uncertainty. The effect should be to downweight the importance of the input data.
- An idea from finance is Titling, which is taking a weighted combination of the the estimator from input data, and the estimator from output data, where the weights are inversely proportional to estimates of the variance of these estimators. This is justified when estimators are independent and normal, and might be a reasonable heuristic otherwise.

How can we combine data with knowledge of system dynamics?

We have model or parameter uncertainty, and some real-world observations $(x, y(x))$ of the quantity $y(x)$ our simulator is simulating, for a few different input conditions x. Our goal is to predict $y(x_*)$ corresponding to some given value x_*.

- Research Question: It seems we should combine our simulation output together with real-world observations for x close to x_*, but how should we do it?
- Research Question: when is it appropriate to ignore the simulation and rely only on real-world data?
- Warning: these questions have been considered, at least partially, by research communities with other kinds of models (e.g., deterministic computer codes).

How can we combine data with knowledge of system dynamics?

Initial thoughts:

- If we are extrapolating (x is far from x_*), it seems that we should rely more on the simulation, and if x_* is close to existing values of x, we should rely more on real-world data.
- In a Bayesian setting, if we could appropriately propagate the parameter and model uncertainty through our model, and if we had a prior distribution on $x \mapsto y(x)$ together with a model for the noise in the real world, we could calculate the posterior distribution of $y(x_*)$.
- In a Bayesian setting with a correctly specified prior and under L^2 loss $(\hat{y}(x_*) - y(x_*))$, including results from our simulation cannot increase our expected loss.
- In the frequentist setting, if our prior may be mis-specified, or if it takes work to specify our prior distribution, we may still wish to ignore the simulation when x_* is close to previous x.

Trace data

If we have trace data, we may want to drive our simulation using it, instead of simulating inputs. This avoids uncertainty about input parameters and input distributions, but brings up many issues.

Research Questions:

- Should we use all of the data, or some subset of the data?
  - For computational reasons, we may not be able to use all of the data.
  - Some of the data may be more "representative" than others, e.g., in finance, where some data is older. How do we choose?
- Research Question: How do we do variance reduction and output analysis? For example, if we use trace data for some of our inputs, but simulate other inputs, can we create variance reduction techniques?
- What if I don't have enough trace data? Can I augment the data I have, e.g., by permuting it?
- What if I have trace data from "inside" the simulation, for example, we have a record of queue size over time for one particular queue in a queueing network?

Uncertainty influences optimization

We are solving $\max_x E[f(x,\omega)]$, where $f(x,\omega)$ is our simulation output, and then implementing the solution in the real world, where our reward is $g(x)$, and $g(x)$ is not exactly equal to $E[f(x,\omega)]$, we have the following issues:

- Group 1: "There is an impression amongst practioners that simulators are 'approximate' and hence optimization through SO is of limited value. This points to a lack of understanding of the utility of SO."

Uncertainty influences optimization

- If we are doing simulation optimization for a system currently in operation at x_0, our uncertainty about $E[f(x,\omega)] - g(x)$ will be small near x_0 and far from x_0.
  - Research Question: what should we do about this?
  - It may make sense to penalize uncertainty in our optimization, which will push our solution toward x_0,
  - Or, we could offer a Pareto frontier over uncertainty and estimated value $E[f(x,\omega)]$.
  - Group 1: "Most of the time, the user just wants a better solution." "More importantly, he wants **evidence** that this solution is better" Can we understand the user's desire in this context?

Uncertainty influences optimization

- Even if $E[f(x,\omega)]$ is an unbiased estimator of $g(x)$, our estimate of the quality of our optimal solution $\max_x E[f(x,\omega)]$ will be a biased estimator of the real value of the solution we implement. "Optimizer's curse", Management Science 2006, Smith & Winkler.
- Empirically, optimization seems to have a tendency to highlight errors in the model. This can be dangerous, but offers opportunities.
  - Research Question: Can we explain why this is true?
  - Research Question: Can this phenomenon be use for checking models, or improving estimates? e.g., by adding a constraint that the optimal solution under the true model should be in a certain region.