A brief survey of uncertainty quantification

Teresa Portone

November 1, 2022

University of Alabama Mathematics Colloquium
• Born and raised in Charlotte, NC
• Musical family
• First-generation PhD
• Major: Math, numerical track
• Minor: Italian, (almost) photography

• Numerical Linear Algebra
• Theory of Probability
• Real Analysis I/II
• Mathematical Statistics
• Numerical Analysis
• Complex Calculus
• VIGRE, Summer 2011
• Mathematical epidemiology project (Malaria)
- REU, Summer 2012
- Numerical model for embryonic tube formation
PhD Computational Science, Engineering, and Mathematics

Thesis “Representing Model-Form Uncertainty from Missing Microstructural Information”

High-velocity “streaks” cause nonlocality in contaminant transport. Traditional upscaled models ignore nonlocality.
• Optimization and Uncertainty Quantification department
  • Interned Fall 2017
  • Joined as staff January 2020
What is uncertainty quantification (UQ)?

My definition:

The science of characterizing, quantifying, and reducing uncertainties in mathematical models.

General references: [1-3]
UQ has taken off in the last couple decades

“Uncertainty quantification is both a new field and one that is as old as the disciplines of probability and statistics.” (Smith 2013) [1]
What is uncertainty?

Inability to assign an exact value to a modeled quantity.

How does uncertainty arise?

**Intrinsic variability** in a modeled quantity

\[
\begin{align*}
fs &= -kx(t) \\
f_d &= -cx'(t)
\end{align*}
\]

**Lack of precise knowledge** of a modeled quantity

\[
\begin{align*}
fs + f_d &\equiv F = ma \equiv m \frac{d^2x(t)}{dt^2} \\
m \frac{d^2x(t)}{dt^2} + c \frac{dx(t)}{dt} + kx(t) &= 0 \\
x(0) &= 1, \quad \frac{dx(t)}{dt}(0) = 0
\end{align*}
\]
Common sources of uncertainty

Model parameters

\[ m \frac{d^2x(t)}{dt^2} + c \frac{dx(t)}{dt} + kx(t) = 0 \]

Boundary + initial conditions

\[ x(0) = 1, \quad \frac{dx(t)}{dt}(0) = 0 \]

Model-form uncertainty

\[ f_s = -kx(t) \]

Measurement uncertainty

\[ f_d = -cx'(t) \]
Result: uncertainty in model predictions

- Model parameters
- Model-form uncertainty
- Boundary + initial conditions
- Measurement uncertainty

Quantity of interest (QoI)
Real-world example

**Quantity of Interest**
- Concentration of radionuclide $^{129}$I in nearby aquifer after $10^6$ years

**Sources of uncertainty**
- Properties of canisters holding nuclear waste (e.g. degradation rate)
- Subsurface properties (e.g. porosity, permeability)
- Environmental conditions (e.g. incidence of earthquake, glaciation)
How do we characterize uncertainty?

Represent sources of uncertainty as random variables (RVs)

Encode what is known through the parameterization of the RV

\[
\begin{align*}
    f_s &= -kx(t) \\
    f_d &= -cx'(t)
\end{align*}
\]

\[
m \sim \mathcal{U}[0.9, 1.1] \quad c \sim \mathcal{N}(0.1, 0.1^2) \quad k \sim \mathcal{N}(2, 0.1^2)
\]

\[
p_{\text{log } \mathcal{N}}(\mu, \sigma^2)(x) \equiv \frac{1}{x \sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2 \sigma^2} \left( \log(x) - \mu \right)^2 \right)
\]
How do we quantify uncertainty?

Propagate sources of uncertainty to QoIs

\[ f_s = -kx(t) \]
\[ f_d = -cx'(t) \]

Compute statistics of QoIs, e.g. mean, variance, tail probabilities
How do we reduce uncertainty?

Use data to gain more precise knowledge of sources of uncertainty

\[
m \frac{d^2x(t)}{dt^2} + c \frac{dx(t)}{dt} + kx(t) = 0
\]

\[d = x_{true}(t = 10) + \epsilon_m, \quad \epsilon_m \sim \mathcal{N}(0, \sigma^2)\]

What is the likelihood the model produced the data for a given \(k, c\)?

\[
\mathcal{M}(k, c) \equiv x(t = 10; k, c)
\]

\[d = \mathcal{M}(k, c) + \epsilon_m, \quad \epsilon_m \sim \mathcal{N}(0, \sigma^2)\]

\[d - \mathcal{M}(k, c) \sim \mathcal{N}(0, \sigma^2)\]

\[
p(d|k, c) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(d - M(k, c))^2}{2\sigma^2}\right)
\]
How do we reduce uncertainty?

Use data to gain more precise knowledge of sources of uncertainty

\[ p(d|k, c) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(d - M(k, c))^2}{2\sigma^2} \right) \]

\[ p(k, c) = p(k)p(c) \]

Bayes' Theorem

\[ p(k, c|d) = \frac{p(d|k, c)p(k, c)}{\int p(d|k, c)p(k, c) dk dc} \]
How do we *reduce* uncertainty?

Use data to gain *more precise knowledge* of sources of uncertainty

This is also called *Bayesian inference* or *Bayesian calibration*
How do we

• characterize
• quantify
• reduce

uncertainty in practice?
Harsh reality

- Models for practical problems challenging
  - Nonlinear: propagating uncertainty + performing inference need many model evaluations
  - Computationally expensive; can afford few evaluations, causing poor statistical accuracy
  - High-dimensional problems

- Data expensive or impossible to attain

- Models imperfect representations of reality
  - Leads to unquantified error in predictions, biased Bayesian calibrations
  - But the correct model form is generally unknown (model-form uncertainty)

- Input/data uncertainties have to be modeled
  - If no quantitative information, have to encode prior belief through expert elicitation [4]
Real-world example

Data acquisition
- Subsurface properties: drill borehole(s)
- Canister properties: fabricate and test several canister specimens in the lab

Computational cost per model evaluation
- ~1.5 hours on 512 cores
- ~8 days on 4 cores
- ~22 years for 1000 samples on 4 cores
Research areas in UQ

- Reduced-order/surrogate models
- Optimal experimental design
- Bayesian inverse problems
- Sensitivity analysis
- Multimodel methods
- Algorithms for high dimensionality
- Model-form uncertainty
Goal: reduce computational burden of UQ by using approximate representations of model.

\[ M(\theta) \approx f(\theta) \]

Common statistical approaches

- Gaussian processes [5,6]
- Stochastic expansions (polynomial chaos, stochastic collocation) [7]

Common reduced-order model (ROM) approaches

- Proper orthogonal decomposition [8,9]
- Principal component analysis [10]

And, more recently

Machine learning!

https://gregorygundersen.com/blog/2019/06/27/gp-regression/
Ongoing opportunities

- Efficient surrogates/ROMs for high-dimensional models
- Adaptive/goal-oriented surrogate/ROM construction
- More theory for error incurred in UQ analyses by using approximation of $M(\theta)$
- Multimodel surrogates
Research areas in UQ

- Efficient forward propagation
- Bayesian inverse problems
- Reduced-order/surrogate models
- Optimal experimental design
- Sensitivity analysis
- Algorithms for high dimensionality
- Model-form uncertainty
Two types of high dimensionality

- Infinite dimensionality
- High cardinality
Ongoing opportunities

• Dimension reduction (PCA [10], active subspaces [11], autoencoders [12], ISOMAP [13])
  • Methods encouraging/exploiting sparsity

• Expanded theory for infinite-dimensional problems with less restrictive assumptions (e.g. linearity, Gaussianity) [14,15]

• Improve on existing inference methods
  • Derivative-based (MALA/HMC, Stochastic Newton, VI) [16-19]
  • Data-informed (DILI) [20]

• Methods to address high cardinality, especially for
  • Surrogates
  • Bayesian inference
  • Sensitivity analysis
Research areas in UQ

- Reduced-order/surrogate models
- Optimal experimental design
- Bayesian inverse problems
- Sensitivity analysis
- Multimodel methods
- Algorithms for high dimensionality
- Model-form uncertainty
Idea: exploit lower-fidelity, cheaper models to lower cost for same accuracy
**Idea:** exploit lower-fidelity, cheaper models to lower cost for same accuracy

**Discretization**

**Modeling assumptions**
**Sampling-based methods**

\[
\hat{M}(\theta) = \frac{1}{N} \sum_{i=1}^{N} M(\theta^{(i)}), \quad \theta^{(i)} \sim p(\theta) \text{ i.i.d.}
\]

\[
\mathbb{V}[\hat{M}] = \frac{\mathbb{V}[M]}{N}
\]
Sampling-based methods – control variates

\[ M_1(\theta) \quad c_1 = \frac{C_1}{C} \ll 1 \quad \text{corr}(M, M_1) = \rho \]

\[ \widehat{M}_{CV}(\theta) = \widehat{M}(\theta) + \alpha(\widehat{M}_1(\theta) - \mathbb{E}[M_1]) \quad \mathbb{E}[\widehat{M}_{CV}(\theta)] = \mathbb{E}[M] \]

\[ \mathbb{V}[\widehat{M}_{CV}(\theta)] = \frac{1}{N} \left( \mathbb{V}[M] + \alpha^2 \mathbb{V}[M_1] + 2\alpha \text{Cov}[M, M_1] \right) \]

\[ \alpha^* = \min_\alpha \mathbb{V}[\widehat{M}_{CV}(\theta)] = -\frac{\text{Cov}[M, M_1]}{\mathbb{V}[M_1]} \]

\[ \mathbb{V}[\widehat{M}_{CV}(\theta)] = \frac{\mathbb{V}[M]}{N}(1 - \rho^2) \]

Unbiased

\[ \rho^2 \approx 1 \rightarrow \text{orders of magnitude reduction in variance} \]
Sampling-based methods – beyond control variates

\[ \hat{M}_{CV}(\theta) = \hat{M}(\theta) + \alpha(\hat{M}_1(\theta) - E[M_1]) \]

Have to estimate this too

Multifidelity Monte Carlo [21]:

\[ \hat{M}_{MFMC} = \hat{M}(\theta) + \alpha(\hat{M}_1(\theta) - \hat{M}_1(\theta_1)) \]

\[ \alpha^* = -\frac{\rho \sqrt{\text{Var}[M]}}{\sqrt{\text{Var}[M_1]}} \]

\[ r_1^* = \sqrt{\frac{\text{Cost}(M)\rho^2}{\text{Cost}(M_1)(1 - \rho^2)}} \]

\[ N_1 = \lceil r_1 N \rceil \]

\[ \text{Var}[\hat{M}_{MFMC}] = \frac{\text{Var}[M]}{N} \left(1 - \rho^2 \left(\frac{r_1 - 1}{r_1}\right)\right) \]
Multimodel methods

- Theory extends to multiple (nonhierarchical) models
- Many algorithms combining different models and sample sets in different ways [21-24]
- Recent focus on multifidelity surrogates, e.g. Gaussian processes [25], multifidelity polynomial chaos [26], and several others [27-29]

\[
M(\theta) \approx f(\theta)
\]

vs.

\[
M_1(\theta) \approx f_1(\theta) \\
M(\theta) - M_1(\theta) \approx f_\Delta(\theta) \\
M(\theta) \approx f_1(\theta) + f_\Delta(\theta)
\]
Multimodel methods: ongoing opportunities

• Moving beyond functions of moments, e.g. tail probabilities, CDFs

• Startup cost of sampling all models to compute sample correlations → exploration vs exploitation tradeoff to find optimal model ensemble

• Stochastic models: stochasticity weakens correlation, but averaging it out can models too costly

• Addressing dissimilar parametrization (high and low fidelity models don’t have same uncertain parameters)
Research areas in UQ

- Reduced-order/surrogate models
- Bayesian inverse problems
- Multimodel methods
- Optimal experimental design
- Sensitivity analysis
- Algorithms for high dimensionality
- Model-form uncertainty
Challenges and ongoing opportunities

\[
p(d|\theta) = \frac{p(d|\theta)p(\theta)}{\int p(d|\theta)p(\theta)d\theta}
\]

\[
p(d|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left( - \frac{(d - M(\theta))^2}{2\sigma^2} \right)
\]

If \( M(\theta) \) nonlinear, can’t compute analytically.

Markov Chain Monte Carlo [14,16-18,19] and Variational Inference [19] methods numerically approximate \( p(\theta|d) \)--need many model evaluations

Much work in multimodel, derivative-based, & surrogate/reduced-order modeling methods to make this tractable. Methods in optimization can be leveraged

Opportunities for improvement: methods addressing multimodal and/or non-Gaussian posteriors; high dimensionality; model error
Research areas in UQ

- Reduced-order/surrogate models
- Optimal experimental design
- Bayesian inverse problems
- Sensitivity analysis
- Multimodel methods
- Algorithms for high dimensionality
- Model-form uncertainty
What model inputs (parameters) most affect model predictions?

Sensitivity analysis methods provide a quantitative measure of output sensitivity to each input [30]

Extremely powerful tool in mathematical modeling. Supports

- **Scientific discovery/model interpretation** – increase understanding of relationships between inputs + their interactions and outputs
- **Dimension reduction** – parameters identified to not affect model predictions can be screened out of further uncertainty analysis
- **Model improvement** – resources can be focused on reducing uncertainties where they will have the most impact
A range of methods

Correlation coefficients

\[ \rho(\theta_i, M(\theta)) = \frac{\text{Cov}[\theta_i, M(\theta)]}{\sqrt{\text{Var}(\theta_i)\text{Var}(M(\theta))}} \]

Estimated from input/output samples

Slope doesn’t matter, just strength of linear relationship

Nonlinear/nonmonotonic dependencies will not be detected.

Higher-order dependencies (i.e. dependence on two parameters varying together) won’t be detected.

Source: http://en.wikipedia.org/wiki/Correlation
Can detect which parameter(s) would be informed in calibration before even collecting data!

\[ f_s = -kx(t) \]

\[ f_d = -cx'(t) \]

\[ x(t = 10) \]

---

\[ \rho = 0.03 \]

\[ \rho = -0.80 \]

---

Probability Density

Prior

Posterior

---

\[ 0.07 \quad 0.10 \quad 0.14 \]

\[ 1.5 \quad 2.0 \quad 2.3 \quad 2.8 \]
A range of methods

Global Variance-Based Sensitivity Analysis [30]

\[ S_i = \frac{\mathbb{V}_{\theta_i} \left[ \mathbb{E}_{\theta \sim i}[M|\theta_i] \right]}{\mathbb{V}[M]} \]

\[ T_i = 1 - \frac{\mathbb{V}_{\theta_i} \left[ \mathbb{E}_{\theta_i}[M|\theta \sim i] \right]}{\mathbb{V}[M]} \]

Effect of varying \( \theta_i \) alone (averaging over other inputs)

Effect of varying \( \theta_i \) alone and with all other inputs

Robust to nonlinearities and higher-order interactions between parameters

# model evaluations: \( N(d + 2) \), \( N \) independent samples, \( d \)-dimensional input space

Assumes inputs statistically independent
A range of methods

• Distribution-based method [31]
  • Instead measure sensitivity of model output distribution.
  • Requires distribution to be estimated—extremely challenging with high input dimension

• Shapley values [32]
  • Game-theory based method
  • Relaxes assumption of independent inputs
  • Computationally costly ($2^d - 1$ evaluations)
Ongoing opportunities [33]

• Computational cost high for more advanced methods, $O(d^\alpha), \alpha \geq 1$

• Computationally tractable methods for correlated inputs

• Unifying process to identify appropriate sensitivity method for a given task/goal
Research areas in UQ

- Efficient forward propagation
- Bayesian inverse problems
- Reduced-order/surrogate models
- Optimal experimental design
- Sensitivity analysis
- Algorithms for high dimensionality
- Model-form uncertainty
Bayesian OED overview

Standard OED problem

\[ d(w) \quad \min_w \Psi(w) = f(p(\theta|d(w))) \]

minimize uncertainty in parameter estimates

Figure courtesy of [34]
Ongoing opportunities

• Outer-loop analysis on expensive Bayesian inverse problem; leverage all efficiency gains possible
  • Surrogates/ROMs
  • Multimodel methods
  • Dimension reduction
  • Derivative-based methods

• Methods to efficiently search experimental design space (especially if it’s high dimensional, e.g. many sensors)

• Methods to address heterogeneous data (i.e. sensor and satellite image data)

• Goal-oriented approaches [35]
Goal-oriented OED overview

\[ d(w) \quad \min_w \Psi^G(w) = f(p(M(\theta)|d(w))) \]

Slide courtesy of Rebekah White

Figure courtesy of [34]
Research areas in UQ

- Efficient forward propagation
- Bayesian inverse problems
- Reduced-order/surrogate models
- Optimal experimental design
- Sensitivity analysis
- Algorithms for high dimensionality
- Model-form uncertainty

Come to my talk tomorrow!
Thanks!

teresaportone.com
References

References


