UQQ

Philip B. Stark

Department of Statistics, UC Berkeley

UQ: Transition Workshop 21–23 May 2012
SAMSI
Research Triangle Park, North Carolina

Abstract

The goal of uncertainty quantification (UQ) is to estimate the uncertainty of models of physical systems calibrated to noisy data (and of predictions from those models), including the contributions of systematic and stochastic measurement error; ignorance; limitations of theoretical models; limitations of numerical representations of those models; limitations of the accuracy and reliability of computations, approximations, and algorithms; and human error (including software bugs). A large portion of UQ research has focused on developing efficient numerical approximations (emulators) of expensive numerical models. In some circles, UQ is nearly synonymous with the study of emulators. I find this unfortunate.

Uncertainty quantification qualification (UQQ) is needed. What sources of uncertainty does a UQ analysis take into account? What does it ignore? How ignorable are the ignored sources? What assumptions were made? What evidence supports those assumptions? Are the assumptions testable? What happens if the assumptions are false? I will sketch a potentially helpful embedding of UQ within the theory of statistical estimation and inverse problems. I will point to a few examples of work that quantifies uncertainty from systematic measurement error and discretization error. Time permitting, I will whine about the 2009 NAS report, “Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile,” and about the misinterpretation of Gaussian process emulators: Both are examples of “unclear proliferation.”

Wisdom that I'll ignore

H.L. Menken

Never argue with a man whose job depends on not being convinced.

Upton Sinclair

It is difficult to get a man to understand something when his salary depends upon his not understanding it.

Unfortunate generalization

It is difficult to get a man to understand something if he *thinks* his salary depends upon his not understanding it.

I hope to convince you that some of the most important issues in UQ are getting the least attention, and that paying attention to them won't hurt your salary.

Of course, I could be wrong on both counts.

What is UQ?

UQ = inverse problems + approximate forward model.

What are inverse problems?

statistics

(so saith Lucien LeCam.)

What is the effect of an approximate forward model, discretization, etc.?

additional systematic measurement errors

How can we deal with systematic measurement errors?

statistics

UQ, again

So, UQ = statistics

(not that I'm chauvinistic; moreover, good applied statistics requires substantive knowledge of the application)

What makes UQ special?

- the particular sources of systematic error
- poorly understood/characterized measurement error
- poorly understood/characterized properties of the underlying "model"
- heavy computational burden (in some applications)
- big data (in some applications)
- heterogeneous and legacy data (in some applications)
- need for speed (in some applications)
- societal consequences (in some applications)

Abstract mumbo-jumbo

How can we embed UQ in the framework of statistics?[1]

Statistical decision theory.
Ingredients:

- *The the state of the world* θ. Math object that represents the physical system.
- Set of possible states of the world Θ. Know *a priori* that $\theta \in \Theta$.
- Observations Y. Sample space of possible observations $\mathcal{Y}$.
- *measurement model* that relates the probability distribution of Y to θ. If η is state of the world, then $Y \sim \mathbb{P}_\eta$. Incorporates the forward model.
- one or more *parameters* of interest, $\lambda = \lambda[\theta]$
- an *estimator* $\hat{\lambda}(Y)$ of the parameter (might be set-valued)
- a risk function that measures the expected loss from estimating $\lambda[\eta]$ by $\hat{\lambda}(Y)$

[1] Moreover, does it help?

How does UQ fit into this framework?

What's $\mathbf{P}_\theta$?

Can think of systematic errors as additional parameters; need bounds on them or can't say much.

Augment θ, Θ to include the systematic errors as parameters.

They are *nuisance parameters*: the distribution of the data depends on them, but they are not of interest.

What’s missing?

- Given η, do we actually know (or can we simulate from) $\mathbf{P}_\eta$?
  That is, do we know the mapping $\eta \to \mathbf{P}_\eta$?
  If not, more unknowns to take into account.
- Usefully constrained sets Θ of possible models.
- Ways of quantifying/bounding the systematic error.
- Ways of assessing the stochastic errors.
- Estimators $\hat{\lambda}$ for $\lambda[\theta]$ in light of the stochastic and systematic errors, Θ, $\eta \to \mathbf{P}_\eta$.

Back to basics: Data quality

Tendency to gloss over data uncertainties:

- ignore systematic errors
- treat all error bars as if they were SDs
- treat all measurement error as Normal (maybe Poisson)
- treat measurement errors as independent
- ignore data reduction steps, normalization, calibration background fits, etc.
- treat inverse of final Hessian of nonlinear LS as if it characterizes the uncertainty.

Data quality: It ain't what we pretend it is

My first eye-opener: helioseismology. Nominal uncertainties didn't even account for numerical instability in the data reduction.
Cold water on beautiful theoretical approaches.

More: Post-Enumeration Survey data from the U.S. Census; online behavior monitoring that has a huge impact on, e.g., Google and Yahoo!'s stock prices.

Most salient: historical nuclear test data used to calibrate numerical models for "Reliable Replacement Warhead."
Instruments gone, people who recorded the data retired, transformations & data reduction mysterious, lots of ±10%: suspicious. What does ±10% mean?

Can't get off the ground

How can you know how well the model should fit the data, if you don't understand the nature and probable/possible/plausible size of the errors in the data?

Theory and Practice

Jan L.A. van de Snepscheut

In theory, there's no difference between theory and practice. But in practice, there is.

Qualitatively quantified version – unknown UQ master

The difference between theory and practice is smaller in theory than it is in practice.

Grappling with Data Quality isn't Sexy

Academics are rewarded for proving hard theorems, doing heroic numerical work (speed or size), making splashy images that get on the cover of Nature, being "first."

Moreover, we fall in love with technology, models, technique, tools.

Digging into data quality, systematic errors, etc., is hard, crucial, unglamorous, and unrewarded.
(OK, so maybe I'm wrong that paying attention won't hurt your salary.)

No UQ in a vacuum!
Can't Q the U without knowing the limitations of the data.

Bad incentives

Gary Larson

Another case of too many scientists and not enough hunchbacks.

John W. Gardner

The society which scorns excellence in plumbing as a humble activity and tolerates shoddiness in philosophy because it is an exalted activity will have neither good plumbing nor good philosophy: neither its pipes nor its theories will hold water.

Emulators and Emulator Errors

Splines, MARS, GP, kriging, etc. Generically, interpolation.

Exacerbate systematic differences between numerical approximation of forward model and true forward model.

Internal measures of accuracy (e.g., cross-validation error, posterior distributions) can be arbitrarily misleading unless there are strong physical constraints on the regularity of the forward model.

Interesting questions:

- What would need to be true for the internal accuracy measure to be accurate?
- How bad could the internal accuracy measure be, given what we really know about the forward model?

How many points does it take to “train” an emulator?

Can lower-bound the number of function evaluations required to ensure that the emulator agrees with the numerical model within ϵ throughout the domain. (Joint work with Jeff Regier.)

Basic idea: can lower-bound the global, isotropic Lipschitz constant by the Lipschitz constant attained by the data used to train the emulator.
Radius of information argument.

To do better than this lower bound, need to know more about the true and numerical forward models.

A priori regularity from the underlying physics?
A posteriori error estimates for numerical PDE?

Sources of randomness in applied statistics

GP emulators treat the forward model as if it is random. Where does the probability come from?

Four main ways probability arises in applications:

- Physics is itself random: thermodynamics, quantum mechanics, big bang
- Experiments in which assignment to treatment is random
- Hypothetical counterfactuals, e.g., p-values for comparing two populations when neither is a random sample
- Pure invention, e.g., GP models for deterministic systems.

In the last two, gotta remember that the probability isn't real—it's supposed.

If the real situation had come from the assumed distribution, stuff follows.

If *not*, it's just made up numbers.

Misleading to treat it as if it necessarily means anything about the actual situation.

What does the analysis tell us?

If UQ gives neither an upper bound nor a lower bound on a sensibly defined measure of uncertainty, what have we learned?

At the very least, need to list what we have and have not taken into account.

Abridged catalog of sources of uncertainty

Broad categories: calibration data, theoretical approximation to the system, numerical approximation of the theoretical approximation in the simulator, interpolation of the simulated results, sampling and testing candidate models, coding errors, inferential techniques

1. error in the calibration data, including noise and systematic error, and assumptions about these
2. approximations in the model, including physics and parametrization
3. finite-precision arithmetic
4. numerical approximations to the approximate physics embodied in the simulator
5. algorithmic errors in the numerical approximation, tuning parameters in the simulations
6. sampling variability in stochastic algorithms and simulations
7. choices of the training points for the interpolator/emulator
8. choices of the interpolator: functional form, tuning parameters, fitting algorithm
9. choice of the measure of agreement between observation and prediction
10. technique actually used to draw conclusions from the emulated output
11. bugs, data transcription errors, faulty proofs, . . .

Examples with inaccurately known forward models and discretization error

Stark (1992) treats a problem in helioseismology in which the forward model is known only approximately; bounds the systematic error that introduces and takes it into account to find confidence sets for a fully infinite-dimensional model; also gives a general framework.

Evans & Stark (2002) give a more general framework.

Stark (2008) discusses generalizing “resolution” to nonlinear problems and problems with systematic errors.

Gagnon-Bartsch & Stark (2012) treat a problem in gravimetry in which the domain is discretized; bound the systematic error that the discretization introduces and take it into account to find confidence sets for a fully infinite-dimensional model.

Intermission

Evaluation Of Quantification Of Margins And Uncertainties Methodology For Assessing And Certifying The Reliability Of The Nuclear Stockpile (EQMU)

Committee on the Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile, 2009.

http://www.nap.edu/openbook.php?record_id=12531&page=R1

Prepare for a rant . . .

Present company excluded

Fundamental Theorem of Physics

Axiom: Anything that comes up in a physics problem is physics.
Lemma: Nobody knows more about physics than physicists.[a]
Theorem: There's no reason for physicists to talk to anybody else to solve physics problems.

[a] Follows from the axiom: Nobody knows more about *anything* than physicists.

Practical consequence

Physicists often re-invent the wheel. It is not always as good as the wheel a mechanic would build.

Some "unsolved" problems–according to EQMU—are solved. But not by physicists.

Who was on the NAS panel?
(Physicists, nuclear physicists, nuclear engineer, shock physicist, senior manager, probabilistic risk assessor, . . . , and one statistician)

Cream of EQMU (p. 25)

Assessment of the accuracy of a computational prediction depends on assessment of model error, which is the difference between the laws of nature and the mathematical equations that are used to model them. Comparison against experiment is the only way to quantify model error and is the only connection between a simulation and reality. . . .

Even if model error can be quantified for a given set of experimental measurements, it is difficult to draw justifiable broad conclusions from the comparison of a finite set of simulations and measurements. . . . it is not clear how to estimate the accuracy of a simulated quantity of interest for an experiment that has not yet been done. . . . In the end there are inherent limits [which] might arise from the paucity of underground nuclear data and the circularity of doing sensitivity studies using the same codes that are to be improved in ways guided by the sensitivity studies.

Example from EQMU (pp. 9–11, 25–6; notation changed)

Device needs voltage V_T to detonate. Detonator applies V_A. “Boom” if $V_A \geq V_T$.

V_T estimated as $\hat{V}_T = 100\,V$, with uncertainty $U_T = 5\,V$.

V_A estimated as $\hat{V}_A = 150\,V$, with uncertainty $U_A = 10\,V$.

Margin $M = 150\,V - 100\,V = 50\,V$.

Total uncertainty $U = U_A + U_T = 10\,V + 5\,V = 15\,V$.

“Confidence ratio” $M/U = 50/15 = 3\frac{1}{3}$.
Magic ratio $M/U = 3$. (EQMU, p. 46)

“If $M/U >> 1$, the degree of confidence that the system will perform as expected should be high. If M/U is not significantly greater than 1, the system needs careful examination.” (EQMU, p. 14)

Scratching the veneer.

Are V_A and/or V_T random? Or simply unknown?

Are $\hat{V}_A$ and $\hat{V}_T$ design parameters? Estimates from data?

Why should U_A and U_T add to give total uncertainty U?

How well are U_A and U_T known?

If U is a bound on the possible error, then have complete confidence if $M > U$: ratio doesn't matter.

If U isn't a bound, what does U mean?

EQMU says:

“Generally [uncertainties] are described by probability distribution functions, not by a simple band of values.”
(EQMU, p. 13)

“An important aspect of [UQ] is to calculate the (output) probability distribution of a given metric and from that distribution to estimate the uncertainty of that metric. The meaning of the confidence ratio (M/U) depends significantly on this definition . . . ”
(EQMU, p. 15)

Vision 1: *U*s are error bars

Suppose V_A and V_T are independent random variables[2] with known means $\hat{V}_A$ and $\hat{V}_T$, respectively.

Suppose $\mathbb{P}\{\hat{V}_A - V_A \leq U_A\} = 90\%$ and $\mathbb{P}\{V_T - \hat{V}_T \leq U_T\} = 90\%$.

What's $\mathbb{P}\{V_A - V_T \geq 0\}$? Can't say, but . . .

Bonferroni's inequality:

$$\mathbb{P}\{\hat{V}_A - V_A \leq U_A \text{ and } V_T - \hat{V}_T \leq U_T\} \geq 80\%.$$

That's a conservative bound. What's the right answer?

[2] Are they random variables? If so, why not dependent?

Vision 2: *U*s are (multiples of) SDs

"…if one knows the type of distribution, it could be very helpful to quantify uncertainties in terms of standard deviations. This approach facilitates meaningful quantitative statements about the likelihood of successful functioning." (EQMU, p. 27)

Does one ever know the type of distribution? Is the SD known to be finite? Can very long tails be ruled out?

Even if so, that's not enough: what's the joint distribution of V_A and V_T?

If V_A and V_T were independent with means $\hat{V}_A$ and $\hat{V}_T$ and SDs U_A and U_T, the SD of $V_A - V_T$ would be $\sqrt{U_A^2 + U_T^2}$, not $U_A + U_T$.

If they are correlated, SD would be $\sqrt{U_A^2 + U_T^2 + 2U_AU_T}$

If *U*s are multiples of SDs, what's the confidence?

Suppose $U = SD(V_A - V_T)$.
What does $M/U = k$ imply about $\mathbb{P}\{V_A > V_T\}$?

Chebychev's inequality:

$$\mathbb{P}\left\{|V_A - V_B - (\hat{V}_A - \hat{V}_B)| \leq kU\right\} \geq 1 - \frac{1}{k^2}.$$

E.g., $k = 3$ gives "confidence" $1 - 1/9 = 88.9\%$.

C.f. typical Gaussian assumption: $k = 3$ gives "confidence"

$$\mathbb{P}\left\{\frac{V_A - V_B - (\hat{V}_A - \hat{V}_B)}{\sigma(V_A - V_T)} \geq 3\right\} \approx 99.9\%.$$

$$88.9\% < 99.9\% < 100\%.$$

Vision 3: one of each

From the description, makes sense that V_T is an unknown parameter, $\hat{V}_T$ is an already-computed estimate of V_T from data, $\hat{V}_A$ is a design parameter, and V_A is a random variable that will be "realized" when the button is pushed.

If so, makes sense that U_T is an "error bar" computed from data.

Either $V_T - \hat{V}_T \leq U_T$ or not: no probability left, only ignorance.

Whether $\hat{V}_A - V_A \leq U_A$ is still a random event; depends on what happens when the button is pushed.

EQMU is careless about what is known, what is estimated, what is uncertain, what is random, etc.
The "toy" lead example is problematic.

Historical error bars

How to make sense of error bars on historical data? Crucial!

Seldom know how the bars were constructed or what they were intended to represent.

Variability in repeated experiments?

Spatial variability (e.g., across-channel variation) within a single experiment?

Instrumental limitation or measurement error?

Hunch? Wish? Prayer? Knee-jerk “it’s 10%?”

Measuring apparatus retired along with institutional memory. Can’t repeat experiments.

Good quote

(EQMU, p. 27, fn 5)

“To the extent (which is considerable) that input uncertainties are epistemic and that probability distribution functions (PDFs) cannot be applied to them, uncertainties in output/integral parameters cannot be described by PDFs.”

And then gibberish ensues.

Bad quotes

(EQMU, p. 21)

“Given sufficient computational resources, the labs can sample from input-parameter distributions to create output-quantity distributions that quantify code sensitivity to input variations.”

“Sampling from the actual high-dimensional input space is not a solved problem.”

“ . . . the machinery does not exist to propagate [discretization errors] and estimate the uncertainties that they generate in output quantities.”

Fallacy

(EQMU, p. 23)

“Analysis shows that 90 percent of the realistic input space (describing possible values of nature’s constants) maps to acceptable performance, while 10 percent maps to failure. This 90 percent is a confidence number . . . we have a 90 percent confidence that all devices will meet requirements and a 10 percent confidence that all will fail to meet requirements.”

Laplace’s principle of insufficient reason: if there’s no reason to think possibilities have different probabilities, assume that the probabilities are equal.

No evidence of difference $\neq$ evidence of no difference.

Example: Gas thermodynamics

Gas of of n non-interacting particles. Each can be in any of r quantum states; possible values of “state vector” equally likely.

1. Maxwell-Boltzman. State vector gives the quantum state of each particle: r^n possible values.

2. Bose-Einstein. State vector gives # particles in each quantum state: $\binom{n+r-1}{n}$ possible values.

3. Fermi-Dirac. State vector gives the # particles in each quantum state, but no two particles can be in the same state: $\binom{r}{n}$ possible values.

Gas thermodynamics, contd.

Maxwell-Boltzman common in probability theory (e.g., “coin gas”), but but describe no known gas.

Bose-Einstein describes bosons, e.g., photons and He^4 atoms.

Fermi-Dirac describes fermions, e.g., electrons and He^3 atoms.

Outcomes can be defined or parametrized in many ways. Not clear which–if any–give equal probabilities.

Principle of Insufficient Reason is insufficient for physics.

Constraints versus prior probabilities

Bayesian machinery (LANL approach) is appealing but can be misleading.

Capturing constraints using priors adds “information” not present in the constraints.

- Why a particular form?
- Why particular values of the parameters?
- What’s the relation between the “error bars” the prior represents and specific choices?

Distributions on states of nature

Bayes' Rule: $\mathbb{P}(B|A) = \mathbb{P}(A|B)\mathbb{P}(B)/\mathbb{P}(A)$.

"Just math."

To have posterior $\mathbb{P}(B|A)$, need prior $\mathbb{P}(B)$.

The prior matters. Where does it come from?

Misinterpretation of LLNL "ensemble of models" approach to UQ: no prior.

Conservation of Rabbits

Freedman's Principle of Conservation of Rabbits

To pull a rabbit from a hat, a rabbit must first be placed in the hat.

The prior puts the rabbit in the hat.

PRA puts many rabbits in the hat.

Bayes/minimax duality: minimax uncertainty is Bayes uncertainty for least favorable prior.[3]

[3] Least favorable $\neq$ "uninformative."

Bounded normal mean

Know that $\theta \in [-\tau, \tau]$.

Observe $Y = \theta + Z$.

$Z \sim N(0, 1)$.

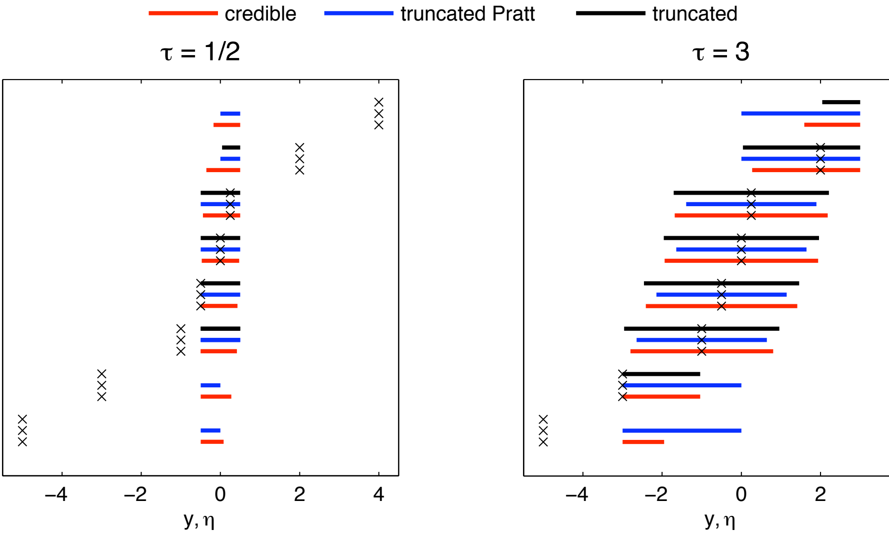
Want to estimate θ.

Bayes: capture constraint using prior, e.g., $\theta \sim U[-\tau, \tau]$.

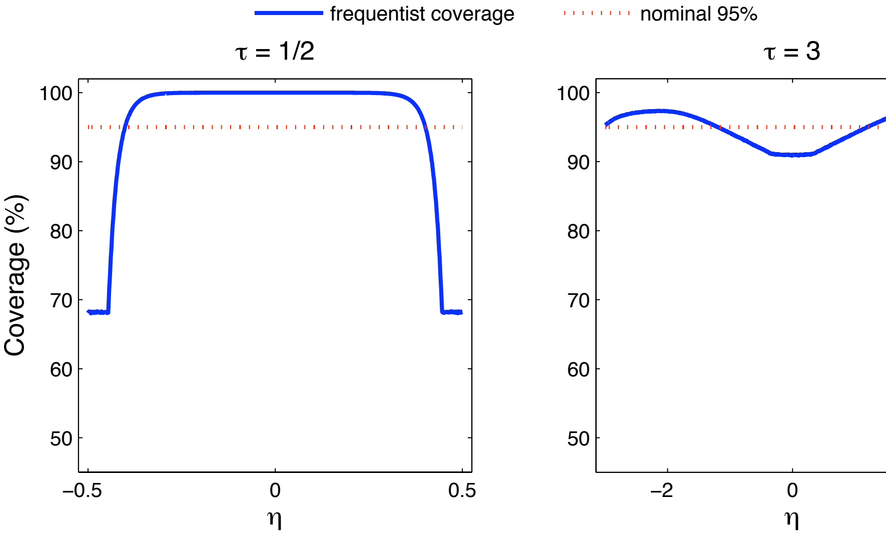
Credible region: 95% posterior probability.

Confidence interval: 95% chance before data are collected.

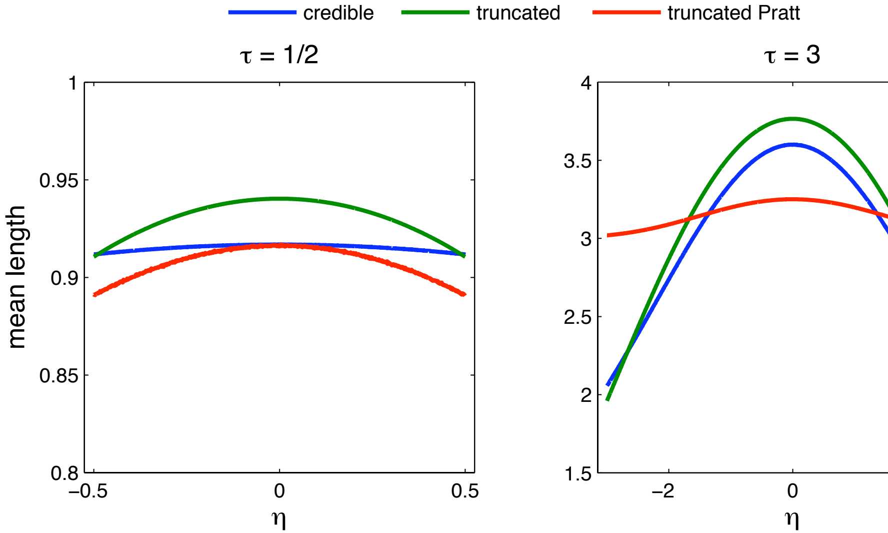
95% Confidence sets vs. credible regions



Coverage of 95% credible regions



Expected size of credible regions and confidence intervals


credible
truncated
truncated Pratt
τ = 1/2
mean length
η
τ = 3
η