Statistics in UQ and UQ in Statistics

Amy Braverman

Jet Propulsion Laboratory, California Institute of Technology

February 26, 2024

Outline

- Problem statement
- Background
- UQ formalism
- Roles of math and stat
- Modifying the UQ formalism
- Model discrepancy
- Methodology
- Summary/discussion

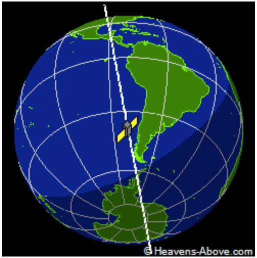
Problem statement

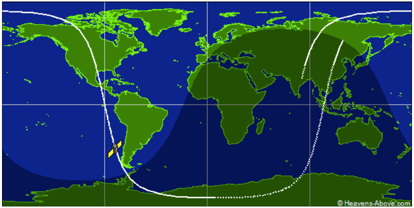
- Develop methods to quantify uncertainty in remote sensing data products delivered by the Orbiting Carbon Observatory 2 and 3 missions.
- The methods must be "off-line" and not interfere with operational data processing.
- They must be computationally efficient enough to keep up with the data stream.
- This problem and our solution are discussed in detail in Braverman et al., 2021. (doi: 10.1137/19M1304283).

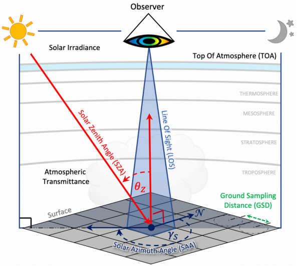
Background

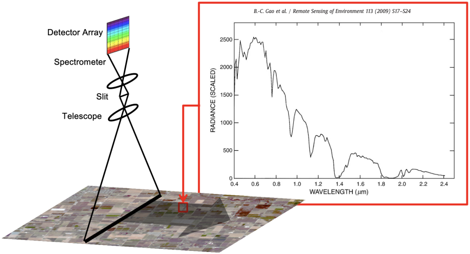
- Passive remote sensing instruments measure photon counts in bins of a discretized electromagnetic spectrum.
- The sun provides incoming photons, which are scattered and absorbed in ways that depend on the media (atmosphere or surface) with which they interact.
- May also be complicated by thermal emission.
- The instrument discretizes the spatial field into “footprints" and aggregates photons over both footprint and spectral bin.

Background


© Heavens-Above.com


© Heavens-Above.com


Observer
Solar Irradiance
Top Of Atmosphere (TOA)
THERMOSPHERE
MESOSPHERE
STRATOSPHERE
TROPOSPHERE
Solar Zenith Angle (SZA)
Line Of Sight (LOS)
Atmospheric
Transmittance
Ground Sampling
Distance (GSD)
Surface
Solar Azimuth Angle (SAA)


Detector Array
Spectrometer
Slit
Telescope
B.-C. Gao et al. / Remote Sensing of Environment 113 (2009) S17–S24
RADIANCE (SCALED)
WAVELENGTH (μm)

Background

Remote sensing levels of data processing:

- Level 0: raw photon counts direct from satellite
- Level 1: georectified and calibrated radiances
- **Level 2: estimates of geophysical state**
- Level 3: “statistical summaries" of Level 2 on uniform space-time grid
- Level 4: output of models or data assimilation

Level 2 “data" aren’t “data"; they are inferences!

When drawing scientific conclusions or making policy decisions, it is crucial to take account of uncertainties in these inferences.

Remote sensing observing system:

True state	Forward function	Noiseless radiance	Instru-ment	Obser-vation	Retrieval	State estimate
$\mathbf{X}$	$\mathbf{F}_0(\cdot, \mathbf{B}_0)$	$\mathbf{Y}_0$	$\mathbf{Y}_0 + \epsilon$	$\mathbf{Y}$	$\mathbf{R}(\cdot, \mathbf{F}_1, \mathbf{B}_1 \ldots)$	$\hat{\mathbf{X}}$

$$\mathbf{X} \longrightarrow \boxed{\mathbf{F}_0(\cdot, \mathbf{B}_0)} \longrightarrow \mathbf{Y}_0 \longrightarrow \boxed{\mathbf{Y}_0 + \epsilon} \longrightarrow \mathbf{Y} \longrightarrow \boxed{\mathbf{R}(\cdot, \mathbf{F}_1, \mathbf{B}_1 \ldots)} \longrightarrow \hat{\mathbf{X}}$$

F_0 = nature's true forward function; $\mathbf{B}_0$ = other true quantities.

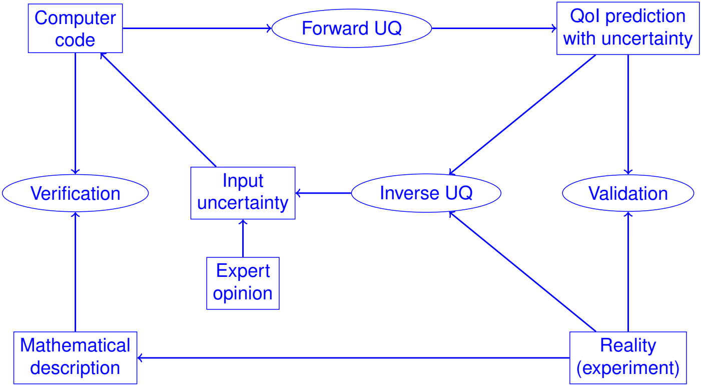
F_1 = forward model used in retrieval, R; $\mathbf{B}_1$ = other retrieval inputs.

ϵ = instrument measurement error.

$\ldots$ = other retrieval algorithm inputs.

UQ formalism

VVUQ:

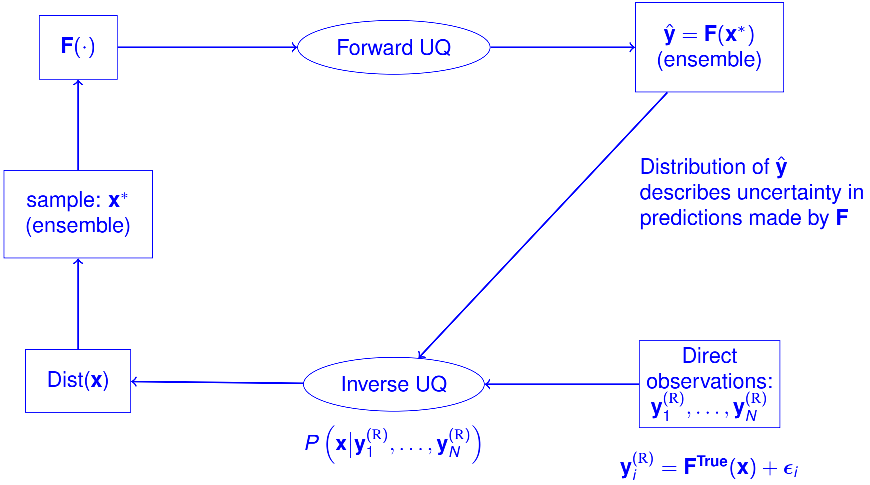


Adapted from Wu et al, (2018). DOI: 10.1016/j.nucengdes.2018.06.004.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

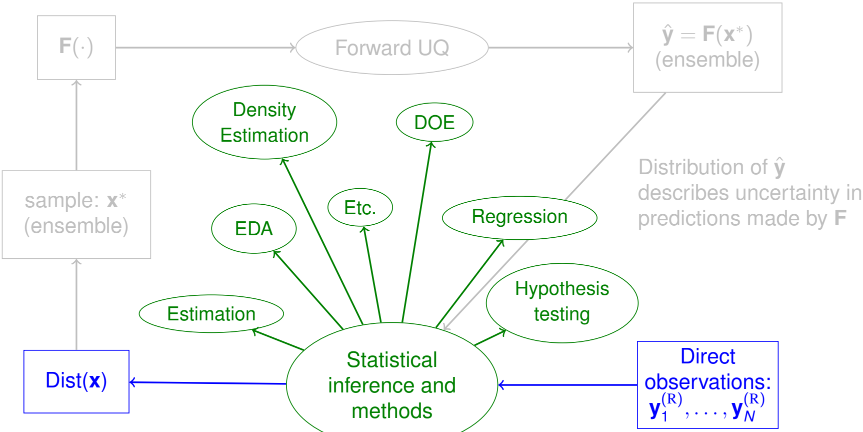
UQ formalism



Roles of Math and Stat

- Statistical methods: inference from observations about unknown probabilistic model
  - estimation and hypothesis testing
  - exploratory data analysis, density estimation, unsupervised learning
  - regression, supervised learning, to uncover significant relationships
  - uncover, test, and quantify relationships from data
  - use estimated model to make statistical predictions with uncertainty.
- Statistical models inherently carry uncertainties with them.

Roles of Math and Stat



$P\left(\mathbf{x}|\mathbf{y}_1^{(R)},\ldots,\mathbf{y}_N^{(R)}\right)$

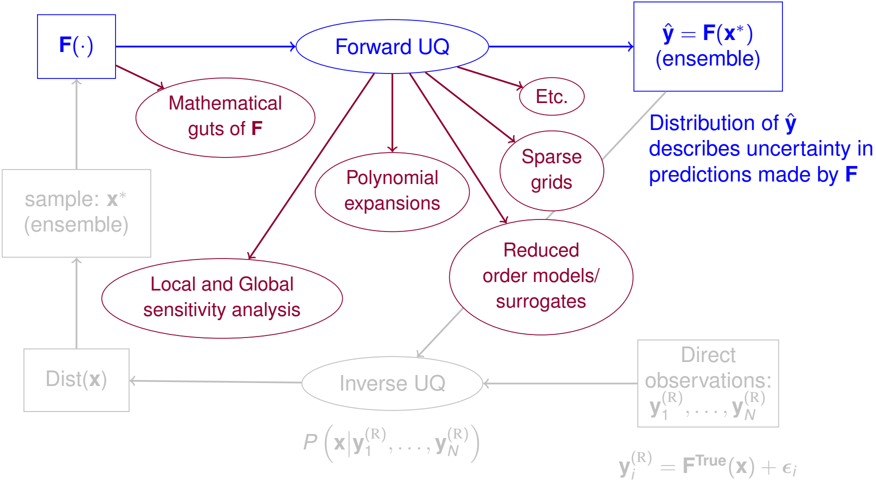
$\mathbf{y}_i^{(R)} = \mathbf{F}^{\mathbf{True}}(\mathbf{x}) + \epsilon_i$

$L\left(\mathbf{x};\mathbf{y}_1^{(R)},\ldots,\mathbf{y}_N^{(R)}\right)$

Roles of Math and Stat

- Mathematical UQ: mathematical approaches for understanding sources of uncertainty in **F** and facilitating efficient forward UQ.
  - exploit structure and properties of **F** to guide forward UQ
  - alternatives to brute-force Monte Carlo forward UQ
  - numerical and other approximations for speed and efficiency
  - optimization!
- Uncertainty expressed through probability distributions, and driven by probabilistic description of input uncertainties.

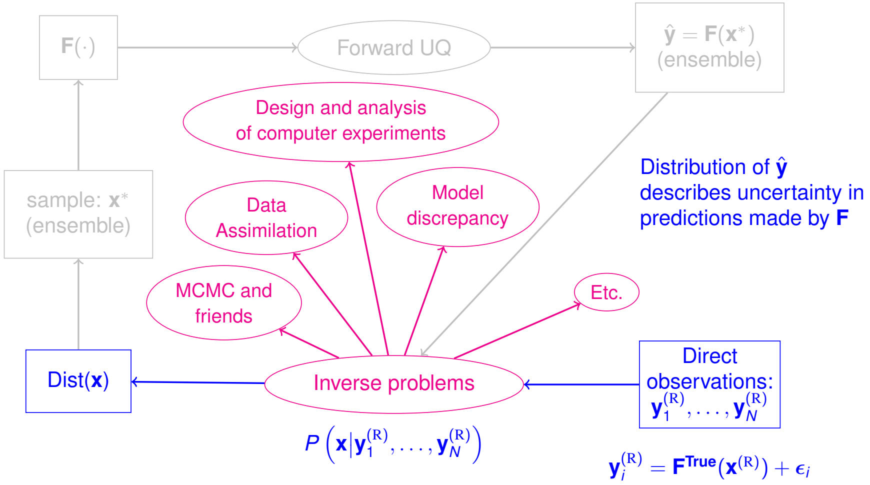
Roles of Math and Stat



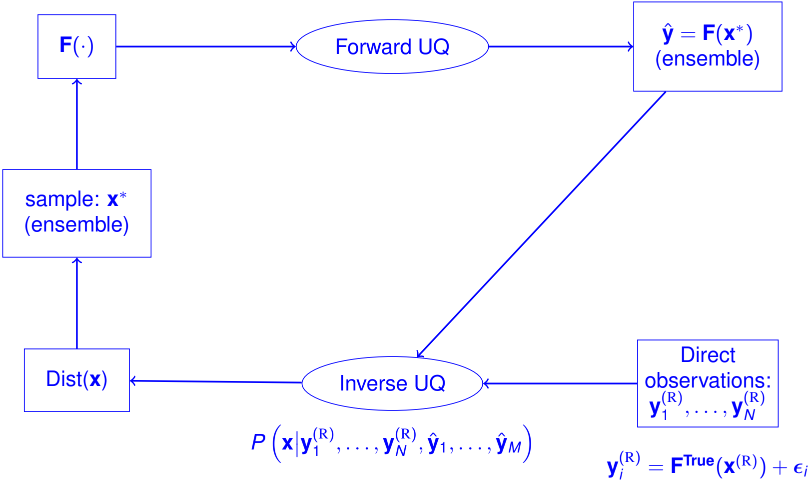
Roles of Math and Stat

- Inverse problems: infer the state of a system from noisy, indirect measurements.
  - heavy use of Bayesian methods
  - overlaps substantially with statistics, but more focussed on this class of problems
  - emphasis on algorithms/samplers
  - because result is a distribution, easy forward propagation

Roles of math and stat

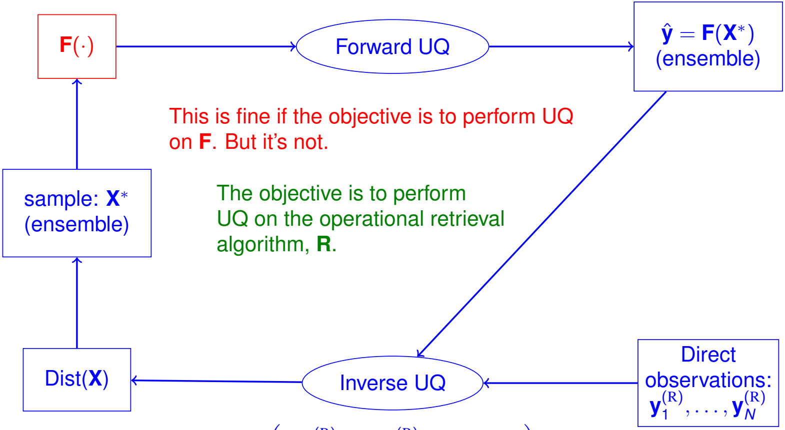


Modifying the UQ formalism



National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Modifying the UQ formalism



This is fine if the objective is to perform UQ on **F**. But it's not.

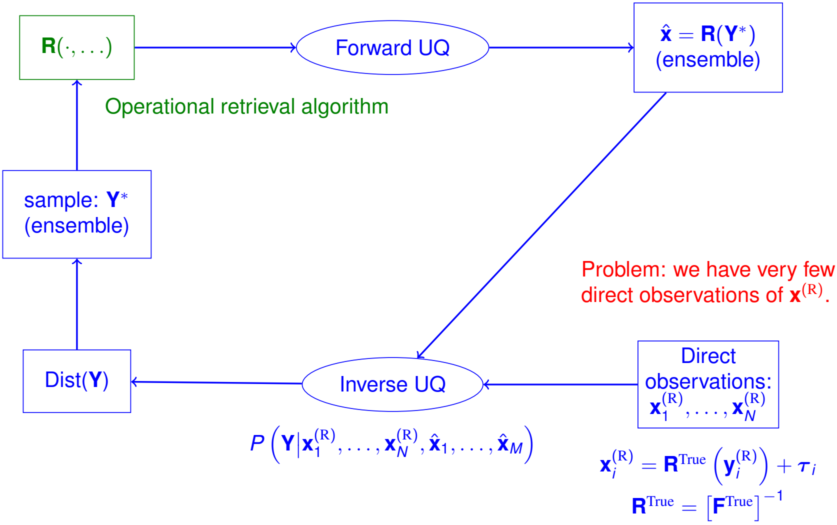
The objective is to perform UQ on the operational retrieval algorithm, **R**.

$$P\left(\mathbf{X} | \mathbf{y}_1^{(\mathrm{R})}, \ldots, \mathbf{y}_N^{(\mathrm{R})}, \hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_M\right)$$

$$\mathbf{y}_i^{(\mathrm{R})} = \mathbf{F}^{\mathrm{True}}\left(\mathbf{x}_i^{(\mathrm{R})}\right) + \epsilon_i$$

Note: this is not the operational retrieval algorithm!

Modifying the UQ formalism

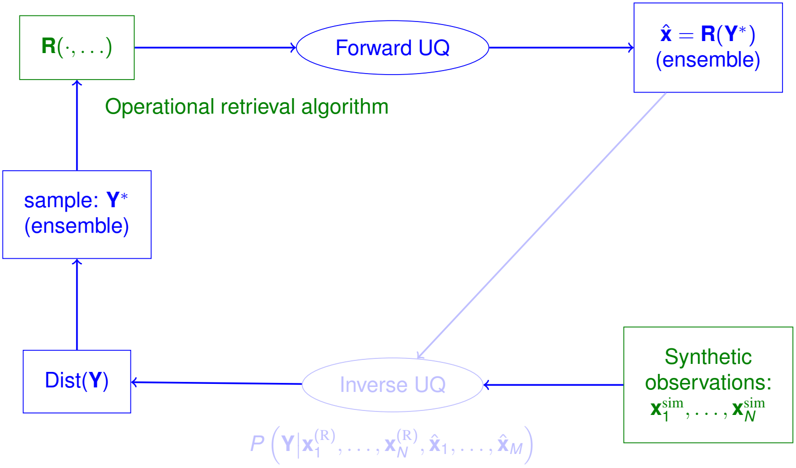


Modifying the UQ formalism

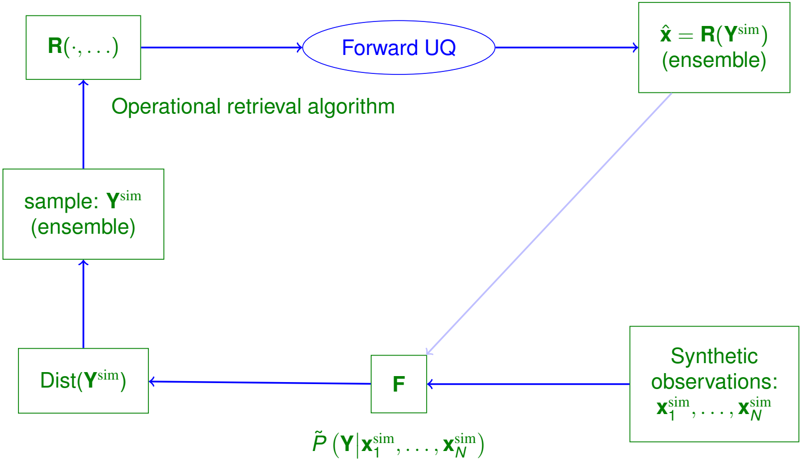
To recap,

- We want to perform UQ on the operational retrieval algorithm, **R**. (Note that **F** is embedded in, and is thus part of, **R**.)
- This requires a computational experiment in which we sample over **R**'s inputs to get an ensemble of outputs ($\hat{\mathbf{x}}$'s) that can be compared to direct observations, $\mathbf{x}^{(\mathrm{R})}$.
- We do not have enough instances of $\mathbf{x}^{(\mathrm{R})}$ to do this.
- Moreover, performing inverse UQ without oversimplifying (e.g., using MCMC) is computationally infeasible. (The oversimplified version *is* **R**).
- So what can we do?

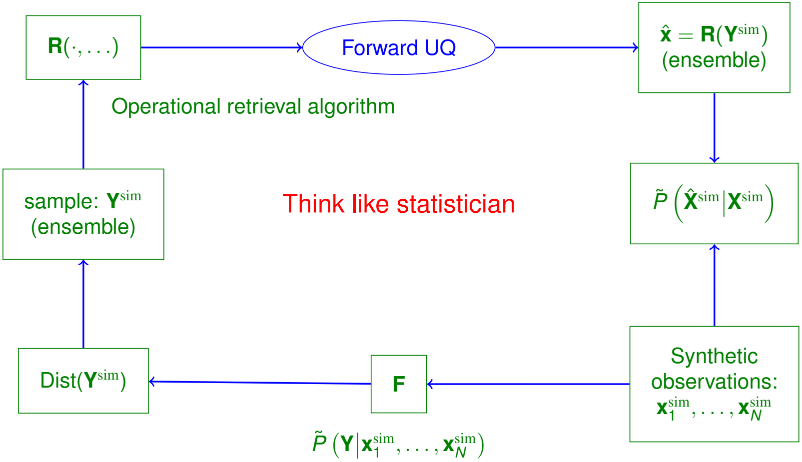
Modifying the UQ formalism



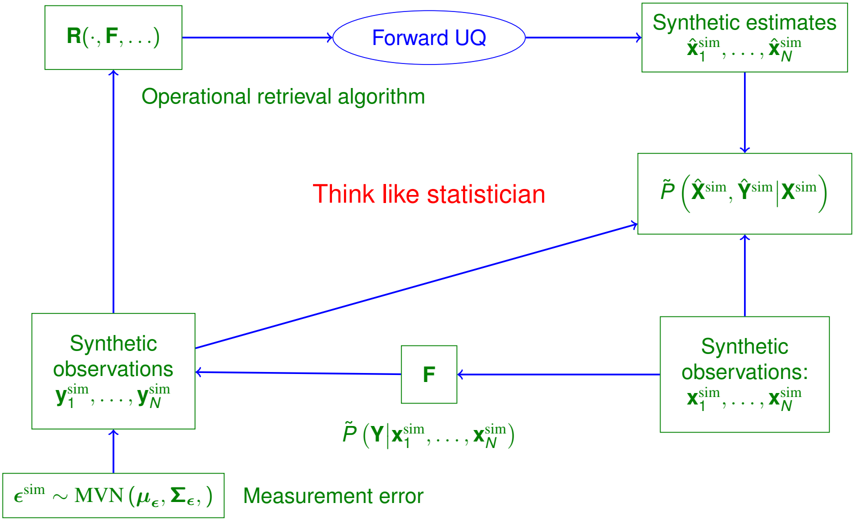
Modifying the UQ formalism



Modifying the UQ formalism



Modifying the UQ formalism



Modifying the UQ formalism

What does "think like a statistician" mean?

- Statisticians invent new estimators and quantify their operating characteristics.
- Here, we quantify the operating characteristics of the system

$$\mathbf{X}^{\mathrm{sim}} \rightarrow \mathbf{F} \rightarrow \mathbf{Y}^{\mathrm{sim}} \rightarrow \mathbf{R} \rightarrow \hat{\mathbf{X}}^{\mathrm{sim}}$$

  with and empirical estimate of $P\left(\hat{\mathbf{X}}^{\mathrm{sim}}, \mathbf{X}^{\mathrm{sim}}\right)$.
- Proposition: uncertainty is quantified by any useful reduction of $\tilde{P}\left(\hat{\mathbf{X}}^{\mathrm{sim}}, \mathbf{X}^{\mathrm{sim}}\right)$, e.g., $\tilde{P}\left(\mathbf{X}^{\mathrm{sim}} | \hat{\mathbf{X}}^{\mathrm{sim}},\right)$.

Modifying the UQ formalism

Fine, but

1. what about the real system?

2. inverse crime: $\mathbf{F}$ is used twice (once to create $\mathbf{y}^{\text{sim}}$ and once in $\mathbf{R}$).

UQ formalism

Fine, but

1. what about the real system?
2. inverse crime: **F** is used twice (once to create $\mathbf{y}^{\text{sim}}$ and once in **R**).

1. We use the learned relationship $\tilde{P}\left(\mathbf{X}^{\text{sim}} \middle| \hat{\mathbf{X}}^{\text{sim}},\right)$ to quantify uncertainty is an actual instance of $\hat{\mathbf{X}}$:

$$\tilde{P}\left(\mathbf{X}^{\text{True}} \middle| \hat{\mathbf{X}}^{\text{Actual}}\right).$$

Modifying the UQ formalism

Fine, but

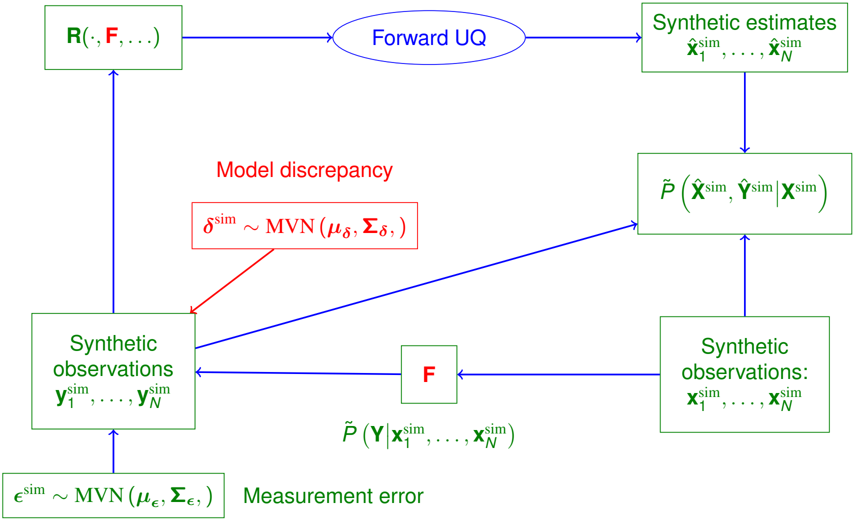
1. what about the real system?
2. inverse crime: **F** is used twice (once to create $\mathbf{y}^{\text{sim}}$ and once in **R**).

1. We use the learned relationship $\tilde{P}\left(\mathbf{X}^{\text{sim}} \middle| \hat{\mathbf{X}}^{\text{sim}},\right)$ to quantify uncertainty in an actual instance of $\hat{\mathbf{X}}$:

$$\tilde{P}\left(\mathbf{X}^{\text{True}} \middle| \hat{\mathbf{X}}^{\text{Actual}}\right).$$

2. Introduce model discrepancy.

Model discrepancy



Model discrepancy

- Model discrepancy is $\delta = \mathbf{F}^{\mathrm{True}}\left(\mathbf{X}^{\mathrm{True}}\right) - \mathbf{F}\left(\mathbf{X}^{\mathrm{True}}\right)$.
- We would like to simulate from the distribution of $\delta \sim \mathrm{MVN}\left(\boldsymbol{\mu}_\delta, \boldsymbol{\Sigma}_\delta\right)$.
- Assume this distribution is Gaussian with mean $\boldsymbol{\mu}_\delta \approx \mathrm{E}(\boldsymbol{\delta}^{\mathrm{sim}})$ and covariance matrix $\boldsymbol{\Sigma}_\delta \approx \mathrm{cov}(\boldsymbol{\delta}^{\mathrm{sim}})$.
- We have noisy samples, $\mathbf{Y}_i = \mathbf{F}^{\mathrm{True}}\left(\mathbf{X}_i^{\mathrm{True}}\right) + \boldsymbol{\epsilon}_i^{\mathrm{True}}, \;\; i = 1, \ldots, N$.
- We don't have $\mathbf{F}\left(\mathbf{X}^{\mathrm{True}}\right)$, but we do have $\left[\mathbf{F}\left(\mathbf{X}^{\mathrm{sim}}\right) - \mathbf{F}(\hat{\mathbf{X}}^{\mathrm{sim}})\right]$, which motivates the approximation,

$$\boldsymbol{\delta}^{\mathrm{sim}} \approx \mathbf{F}^{\mathrm{True}}\left(\mathbf{X}^{\mathrm{True}}\right) - \mathbf{F}\left(\hat{\mathbf{X}}^{\mathrm{Actual}}\right) - \left[\mathbf{F}(\mathbf{X}^{\mathrm{sim}}) - \mathbf{F}(\hat{\mathbf{X}}^{\mathrm{sim}})\right].$$

Model discrepancy

- Let $\mathbf{Y}^{\text{Actual}} \equiv \mathbf{F}^{\text{True}}\left(\mathbf{X}^{\text{True}}\right) + \epsilon^{\text{True}}$, and $\hat{\mathbf{Y}}^{\text{Actual}} \equiv \mathbf{F}\left(\hat{\mathbf{X}}^{\text{Actual}}\right)$, and similarly for simulated. Then,

$$\delta^{\text{sim}} \approx \left(\mathbf{Y}^{\text{Actual}} - \epsilon^{\text{True}} - \hat{\mathbf{Y}}^{\text{Actual}}\right) - \left(\mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right),$$

$$\delta^{\text{sim}} + \epsilon \approx \left(\mathbf{Y}^{\text{Actual}} - \hat{\mathbf{Y}}^{\text{Actual}}\right) - \left(\mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right).$$

- Expected value:

$$\mathrm{E}(\delta^{\text{sim}} + \epsilon^{\text{Actual}}) \approx \mathrm{E}\left(\mathbf{Y}^{\text{Actual}} - \hat{\mathbf{Y}}^{\text{Actual}}\right) - \mathrm{E}\left(\mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right),$$

$$\tilde{\mu}_\delta + \mathbf{0} \approx \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{Y}_n^{\text{Actual}} - \hat{\mathbf{Y}}_n^{\text{Actual}}\right) - \frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{Y}_m^{\text{sim}} - \hat{\mathbf{Y}}_m^{\text{sim}}\right),$$

where $n = 1, \ldots, N$ indexes actual retrievals, and $m = 1, \ldots, M$ indexes trials of the simulation.

Model discrepancy

- Covariance:

$$\operatorname{cov}(\boldsymbol{\delta}^{\text{sim}} + \boldsymbol{\epsilon}^{\text{Actual}}) \approx \operatorname{cov}\left(\mathbf{Y}^{\text{Actual}} - \hat{\mathbf{Y}}^{\text{Actual}}\right) + \operatorname{cov}\left(\mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right) - 2\operatorname{cov}\left(\mathbf{Y} - \hat{\mathbf{Y}}, \mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right)$$

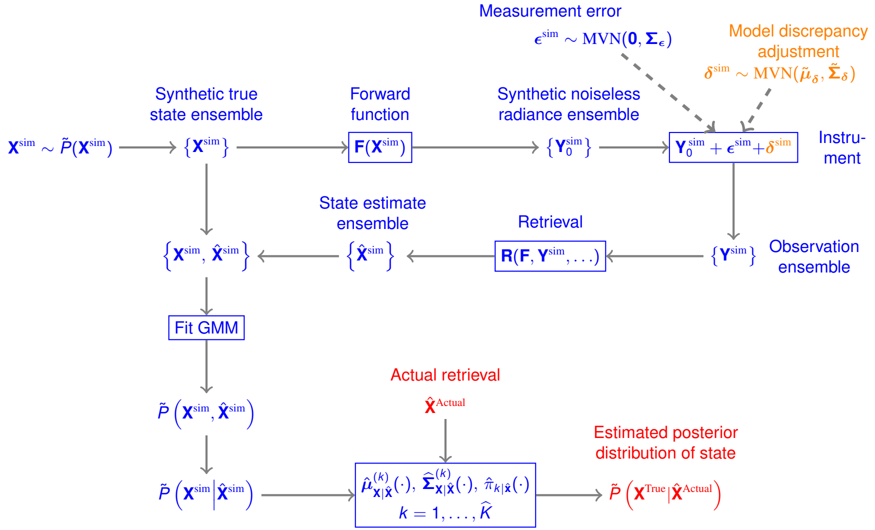
$$\tilde{\boldsymbol{\Sigma}}_{\delta} = \operatorname{cov}\left(\boldsymbol{\delta}^{\text{sim}}\right) \leq \operatorname{cov}\left(\mathbf{Y}^{\text{Actual}} - \hat{\mathbf{Y}}^{\text{Actual}}\right) + \operatorname{cov}\left(\mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right) - \operatorname{cov}(\boldsymbol{\epsilon}^{\text{Actual}}),$$
$$\approx \widehat{\operatorname{cov}}\left(\mathbf{Y}^{\text{Actual}} - \hat{\mathbf{Y}}^{\text{Actual}}\right) + \widehat{\operatorname{cov}}\left(\mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right) - \operatorname{cov}(\boldsymbol{\epsilon}^{\text{Actual}}),$$

assuming $\operatorname{cov}\left(\mathbf{Y}^{\text{Actual}} - \hat{\mathbf{Y}}^{\text{Actual}}, \mathbf{Y}^{\text{sim}} - \hat{\mathbf{Y}}^{\text{sim}}\right) \geq 0$, and $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}^{\text{sim}}$ are independent.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Methodology



Summary/discussion

- UQ community traditionally focusses on uncertainties of deterministic models' output.
- Stat community traditionally focusses on building statistical models, which carry uncertainty with them, but do not explicitly encode mechanistic knowledge.
- Remote sensing is a good example of a problem that combines elements of both.
- Other examples?

Summary/discussion

If the UQ community (as presently constituted) is to expand towards more statistics and statisticians, what parts of that audience should we target?

- design and analysis of computer experiments, and experimental design in general (ASA's UQ Interest Group)
- spatial/spatio-temporal statistics uses GP's and other models with spatial location/time as inputs; leverage this in more general UQ settings (e.g., emulators)
- machine learning is often inherently statistical but uncertainty not emphasized- could we do more?
- inverse problems community that intersects with UQ is not well-represented in mainstream statistics- another opportunity?

Summary/discussion

- Mainstream statisticians contemplate a wider range of applications in which computational models are not necessarily the focus, and exist along side data collected for other purposes.
- Expanding UQ territory will be accomplished by young researchers willing to think “outside the box". It will require a cultural shift.

Acknowledgments

©2024 California Institute of Technology. Government sponsorship acknowledged.