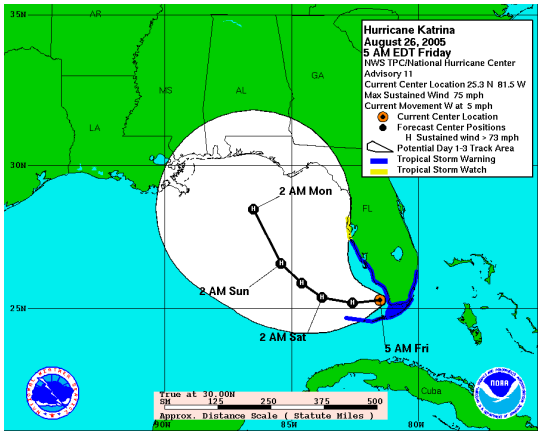
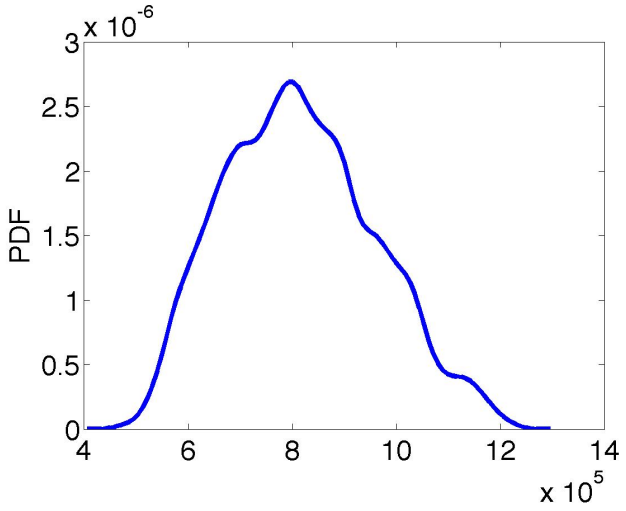
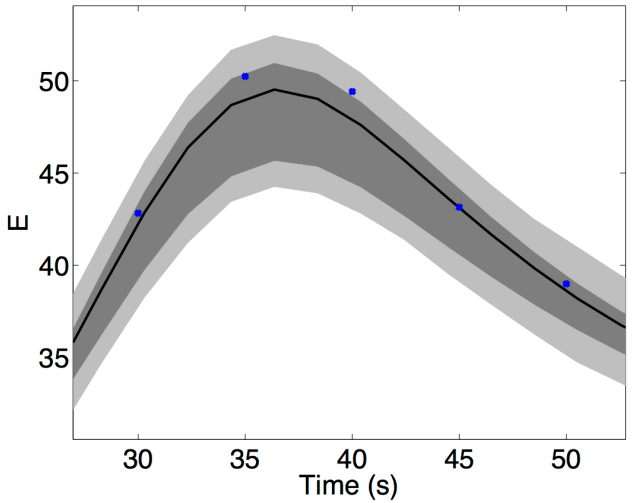
Uncertainty Propagation for Biological Models

Ralph C. Smith
Department of Mathematics
North Carolina State University







Essentially, all models are wrong, but some are useful, George E.P. Box, Industrial Statistician.

Support: DOE Consortium for Advanced Simulation of LWR (CASL)
NNSA Consortium for Nonproliferation Enabling Capabilities (CNEC)
NSF Data-Enabled Science and Engineering of Atomic Structure (SEAS)
NSF Collaborative Research CDS&E
Air Force Office of Scientific Research (AFOSR)

Example 1: HIV Model for Characterization and Control Regimes

HIV Model:

$$\dot{T}_1 = \lambda_1 - d_1 T_1 - (1-\varepsilon) k_1 V T_1$$

$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1-f\varepsilon) k_2 V T_2$$

$$\dot{T}_1^* = (1-\varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1-f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

$$\dot{V} = N_T \delta (T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1-f\varepsilon)\rho_2 k_2 T_2] V$$

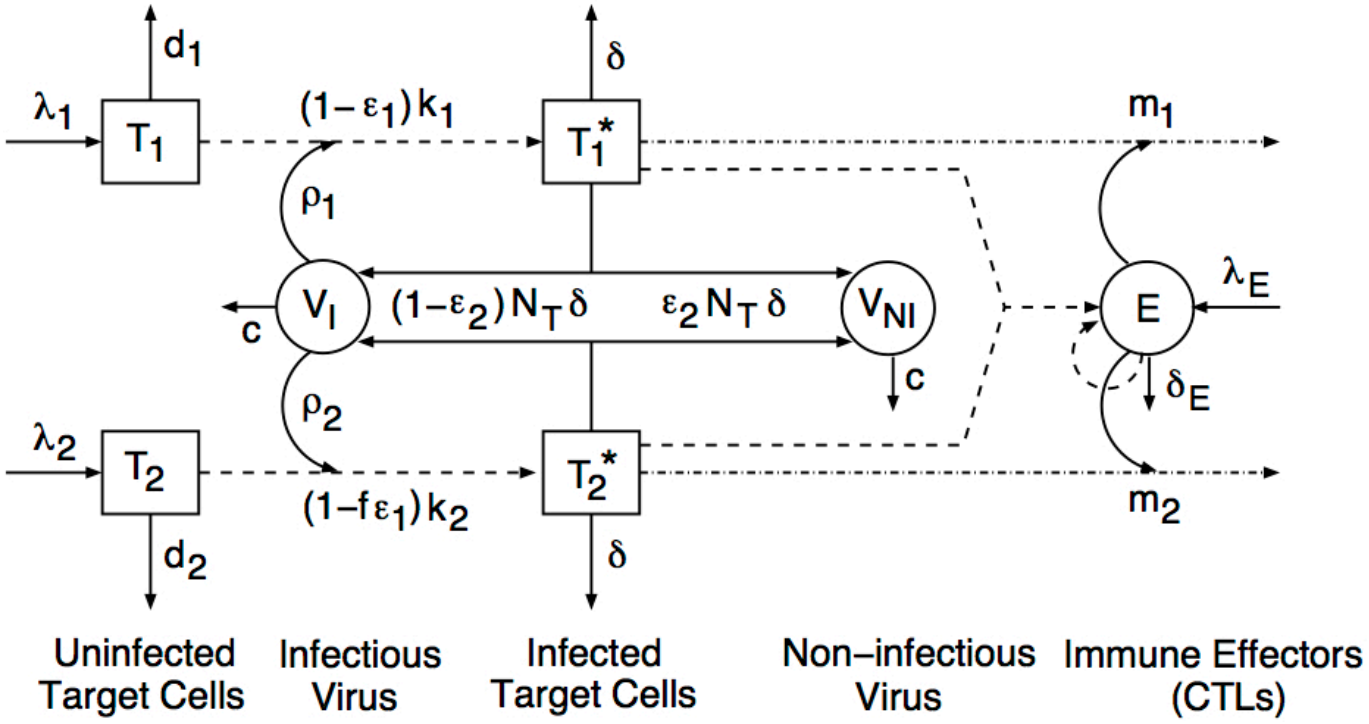
$$\dot{E} = \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E$$

Notes: 21 parameters

[Adams, Banks et al., 2005, 2007]

Notation: $\dot{E} \equiv \frac{dE}{dt}$

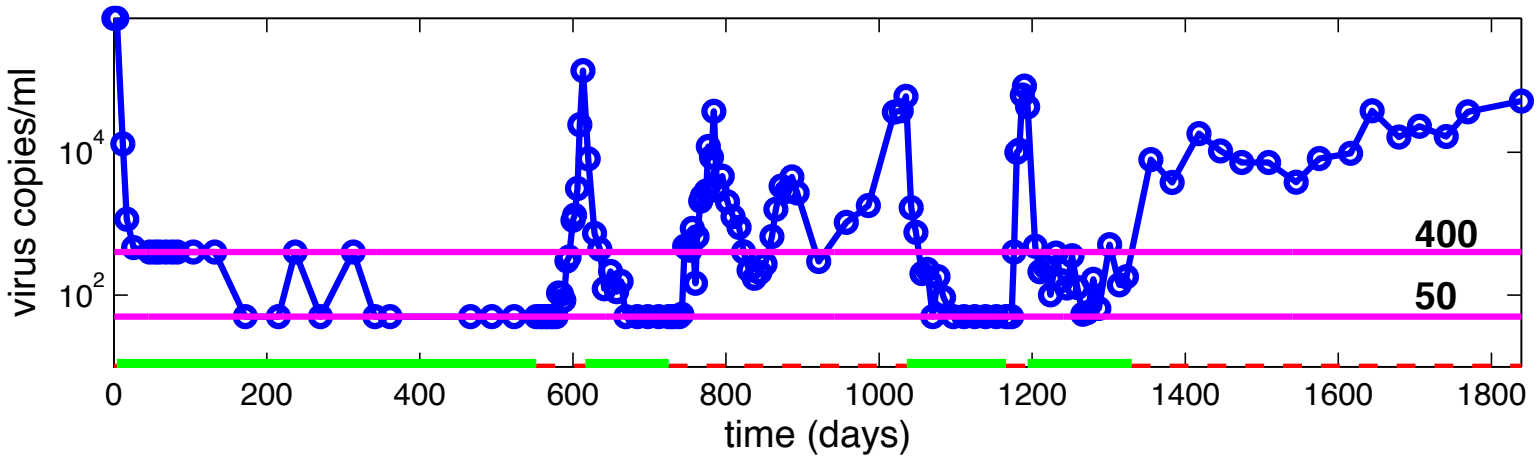
Compartments:



Example: HIV Model for Characterization and Treatment Regimes

HIV Model: Several sources of uncertainty including viral measurement techniques

Example: Upper and lower limits to assay sensitivity



UQ Questions:

- What are the uncertainties in parameters that cannot be directly measured?
- What is optimal treatment regime that is "safe" for patient?
- What is expected viral load? Issue: very often requires high-dimensional integration!
- e.g., $\mathbb{E}[V(t)] = \int_{\mathbb{R}^{21}} V(t,q)\rho(q)dq$

Experimental results are believed by everyone, except for the person who ran the experiment, source anonymous, quoted by Max Gunzburger, Florida State University.

Model Calibration and Uncertainty Propagation

Sources of Uncertainty:

- Model
- Parameters
- Sensor measurements
- Initial conditions

Strategy:

- Quantify uncertainty in parameters
- Propagate uncertainty through model

Example: HIV model

$$\dot{T}_1 = \lambda_1 - d_1 T_1 - (1-\varepsilon) k_1 V T_1$$

$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1-f\varepsilon) k_2 V T_2$$

$$\dot{T}_1^* = (1-\varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1-f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

$$\dot{V} = N_T \delta (T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1-f\varepsilon)\rho_2 k_2 T_2] V$$

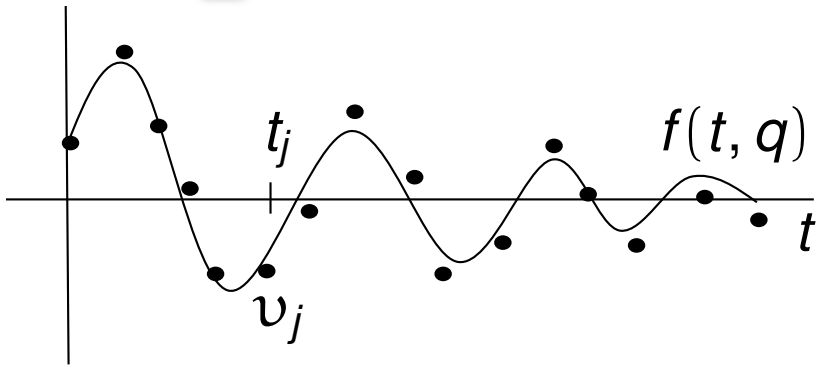
$$\dot{E} = \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E$$

Parameters: Reduced set

$$q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$$

Point Estimates: Ordinary least squares

$$q^0 = \arg\min_q \frac{1}{2} \sum_{j=1}^{N} [\upsilon_j - f(t_j, q)]^2$$



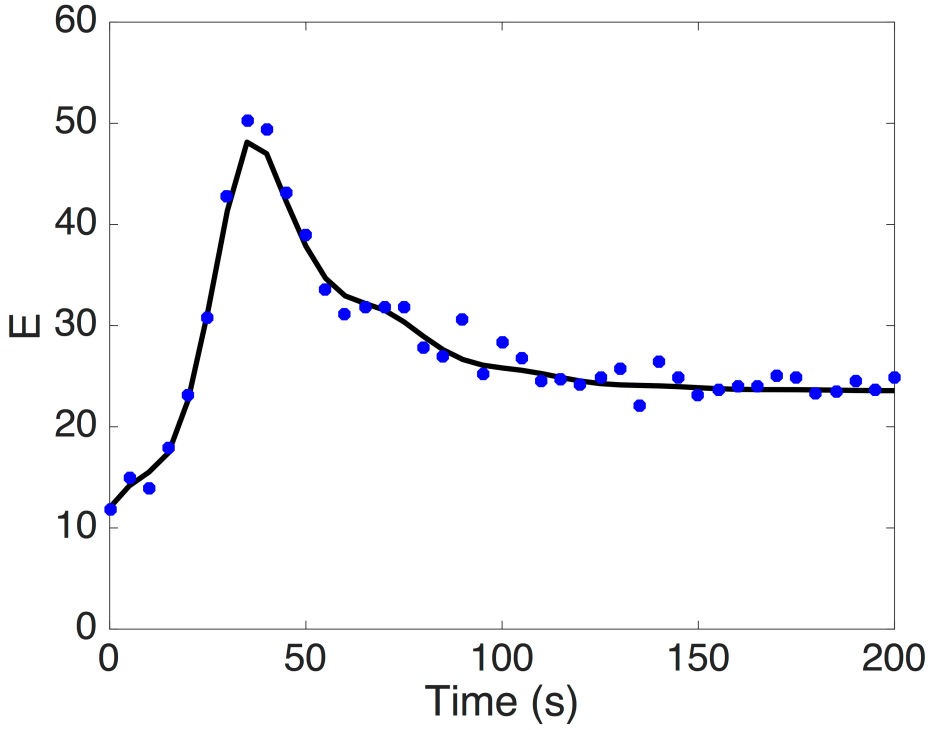
Note: Scaling critical since parameter values vary by 8 orders of magnitude.

Model Calibration and Predictions

Optimization Results:

b_E	δ	d_1	k_2	λ_1	K_b
0.30	0.68	9.1×10^{-3}	1.22×10^{-4}	9.95×10^{3}	88.5

Data and Prediction of Immune Effector Response E:



Note: Point estimates but no quantification of uncertainty in:

- Model
- Parameters
- Data

Goals:

- Replace point estimates with distributions.
- Construct credible and prediction intervals.
- Natural in a Bayesian framework

Example 2: Sinko-Streifer Model

Motivation: Consider use of mosquitofish Gambusia Affinis to control mosquitos. Modeling required to answer the following questions:

- How many fish should be stocked in each paddy?
- How should the fish be initially stocked? All at once or periodically?
- How should they be stocked to augment control already provided by endemic fish without highly damaging the local fish populations?

Example: Sinko-Streifer Model

Notation and Assumptions:

- $u(t,x)$: Number of fish of size x at time t
- μ: Death rate
- Growth rate of same-sized individuals is same and denoted by $g(t,x)$

$$\frac{dx}{dt} = g(t,x)$$

Flux Balance: $u(t,x)$ is a "density" and rate is $\phi(t,x) = g(t,x)u(t,x)$

$$\frac{\partial u}{\partial t} + \frac{\partial \phi}{\partial x} = \text{- deaths} \Rightarrow \frac{\partial u}{\partial t} + \frac{\partial (gu)}{\partial x} = -\mu u$$

Model:

$$\frac{\partial u}{\partial t} + \frac{\partial (gu)}{\partial x} = -\mu u$$

$$g(t,x)u(t,x)\big|_{x=x_0} = \int_{x_0}^{x_1} k(t,\xi)u(t,\xi)d\xi$$

$$u(0,x) = \Phi(x)$$

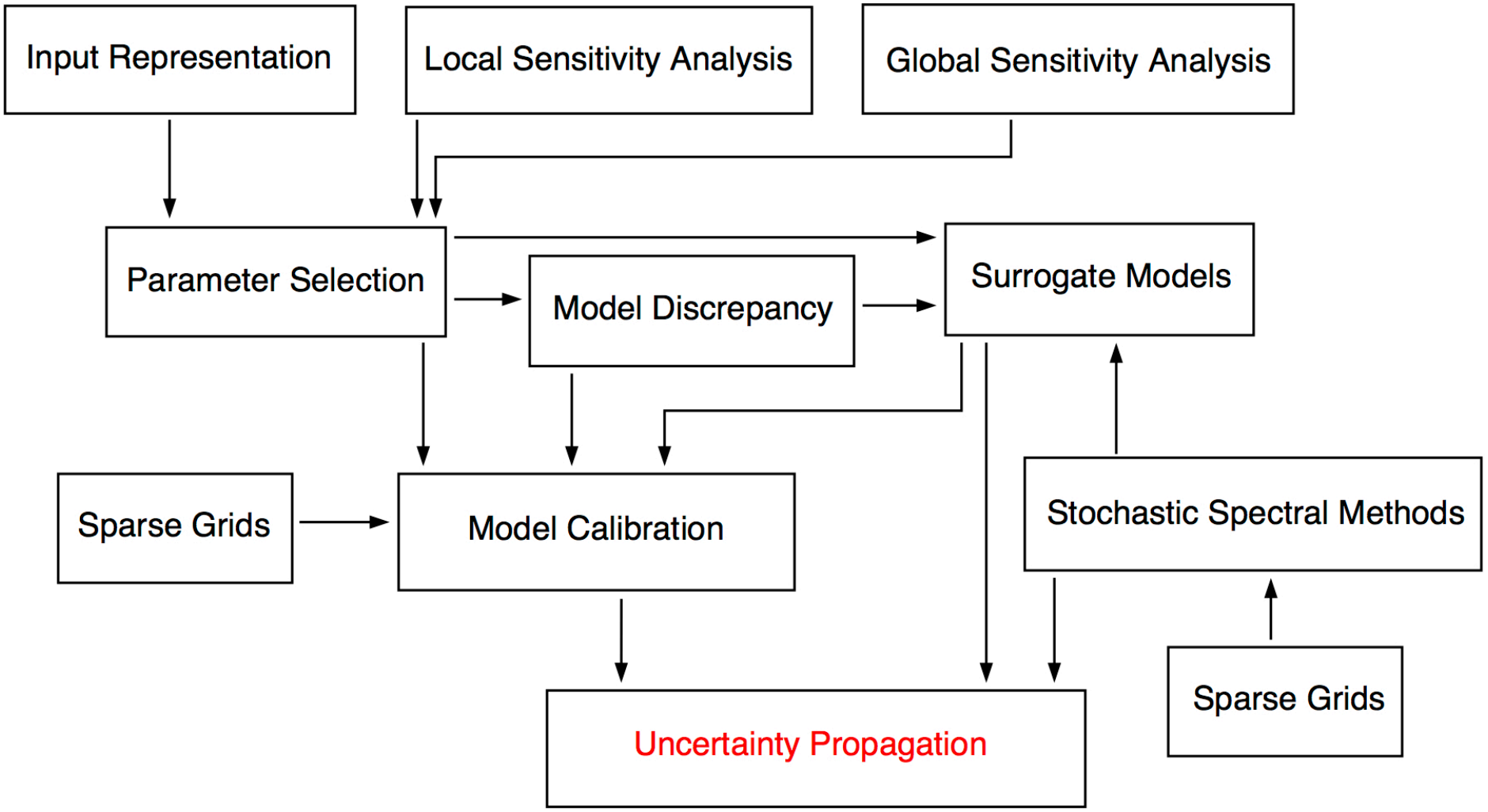
UQ Issues:

- How do we represent $g(t,x)$ or $g(x)$
- How do we propagate uncertainties through PDE?
- How do we construct surrogate models?

Later Talk: Dirk Husmeier

Steps in Uncertainty Quantification

Note: Uncertainty quantification requires synergy between statistics, mathematics and application area.



Uncertainty Propagation

Setting:

- We assume that we have determined distributions for parameters
  - e.g., Bayesian inference (Brian Reich), prior experiments, expert opinion

$$\dot{T}_1 = \lambda_1 - d_1 T_1 - (1-\varepsilon) k_1 V T_1$$

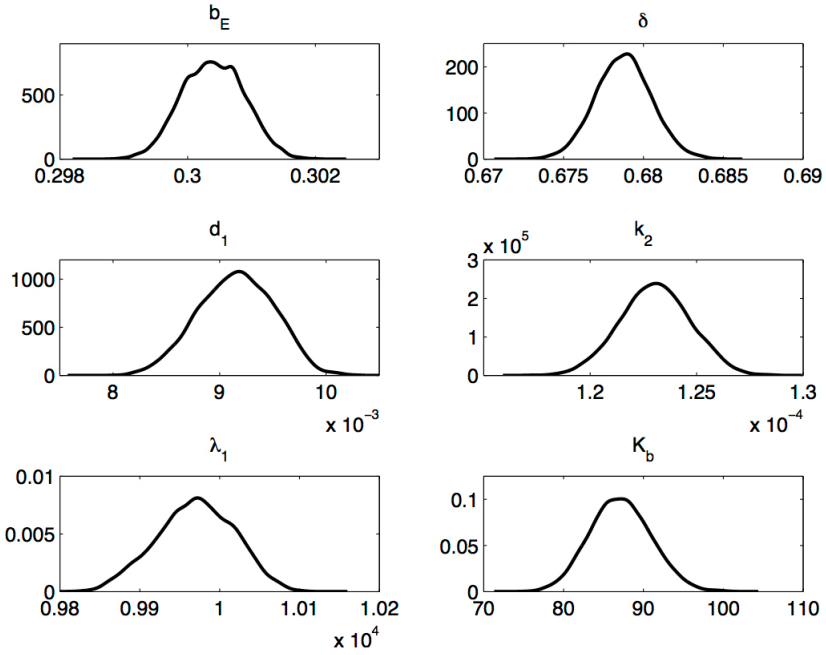
$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1-f\varepsilon) k_2 V T_2$$

$$\dot{T}_1^* = (1-\varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1-f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

$$\dot{V} = N_T \delta (T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1-f\varepsilon)\rho_2 k_2 T_2] V$$

$$\dot{E} = \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E$$



Goal: Construct statistics for quantities of interest

- e.g., Expected viral load in HIV patient with appropriate uncertainty intervals

- Note: Often involves moderate to high-dimensional integration

$$\mathbb{E}[V(t)] = \int_{\mathbb{R}^6} V(t,q)\rho(q)dq$$

Uncertainty Propagation

Issues:

• Uncertainty propagation and computation of statistical quantities of interest much more difficult for PDE models.

• e.g. Sinko-Streifer model: $u(t,x)$: Number of fish of size x at time t

$$\frac{\partial u}{\partial t} + \frac{\partial (gu)}{\partial x} = -\mu u$$

$$g(t,x)u(t,x)\big|_{x=x_0} = \int_{x_0}^{x_1} k(t,\xi)u(t,\xi)d\xi$$

$$u(0,x) = \Phi(x)$$

Random Field Representation:

$$g(x) = \sum_{j=1}^{p} q_j \phi_j(x)$$

Quantity of Interest:

$$\mathbb{E}[u(t,x)] = \int_{\mathbb{R}^p} u(t,x,q)\rho(q)dq$$

Issues:

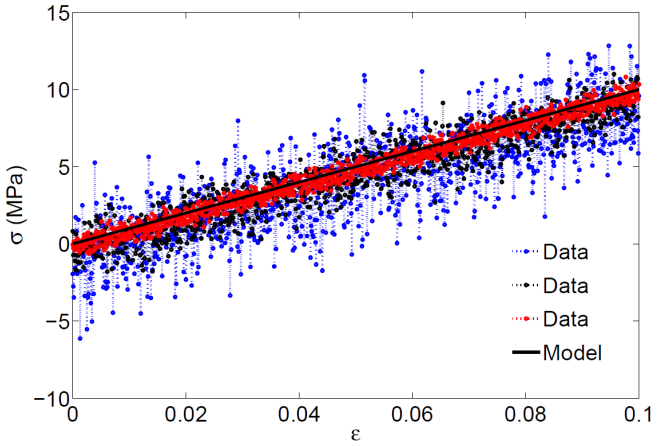
- How do we efficiently propagate input uncertainties through models? Surrogate models.
- How do we approximately integrate in moderate to high dimensions; e.g., p = 10-60? Monte Carlo sampling, sparse grid quadrature

Forward Uncertainty Propagation: Linear Models

Linear Models: Analytic mean and variance relations

Example: Linear stress-strain relation

$$\Upsilon_i = Ee_i + E_2 e_i^3 + \varepsilon_i \ , \ i = 1, \ldots, n$$



Model Statistics:

Let $\overline{E}, \overline{E}_2$ and $\mathrm{var}(E), \mathrm{var}(E_2)$ denote parameter means and variance. Then

$$\mathbb{E}[Ee_i + E_2 e_i^3] = \overline{E} e_i + \overline{E}_2 e_i^3$$

$$\mathrm{var}[Ee_i + E_2 e_i^3] = e_i^2 \mathrm{var}(E) + e_i^6 \mathrm{var}(E_2) + 2e_i^4 \mathrm{cov}(E, E_2)$$

Response Statistics: Assume measurement errors uncorrelated from model response.

$$\mathbb{E}[\Upsilon_i] = \overline{E} e_i + \overline{E}_2 e_i^3$$

$$\mathrm{var}[\Upsilon_i] = e_i^2 \mathrm{var}(E) + e_i^6 \mathrm{var}(E_2) + 2e_i^4 \mathrm{cov}(E, E_2) + \mathrm{var}(\varepsilon_i)$$

Problem: Models are almost always nonlinearly parameterized

Forward Uncertainty Propagation: Sampling Methods

Strategy 1: Randomly sample from parameter and measurement error distributions and propagate through model to quantify response uncertainty.

Advantages:

- Applicable to nonlinear models.
- Parameters can be correlated and non-Gaussian.
- Straight-forward to apply and convergence rate is independent of number of parameters.
- Can directly incorporate both parameter and measurement uncertainties.

Disadvantages:

- Very slow convergence rate: $\mathcal{O}(1/\sqrt{M})$ where M is the number of samples.
- 100-fold more evaluations required to gain additional place of accuracy.
- This motivates numerical analysis techniques.

Uncertainty Propagation

Sampling-Based Approaches:

- Quadrature: Monte Carlo, Latin hypercube, Sobol'
- Interval definitions and construction
- Prediction intervals for HIV model via DRAM algorithm

Numerical Analysis-Based Approaches:

- Stochastic Galerkin, stochastic collocation, discrete projection
- Regression-based methods with sparsity control (LASSO)

Numerical Quadrature

Motivation: Computation of expected values requires approximation of integrals

$$\mathbb{E}[u(t,x)] = \int_{\mathbb{R}^p} u(t,x,q)\rho(q)dq$$

Example: HIV model

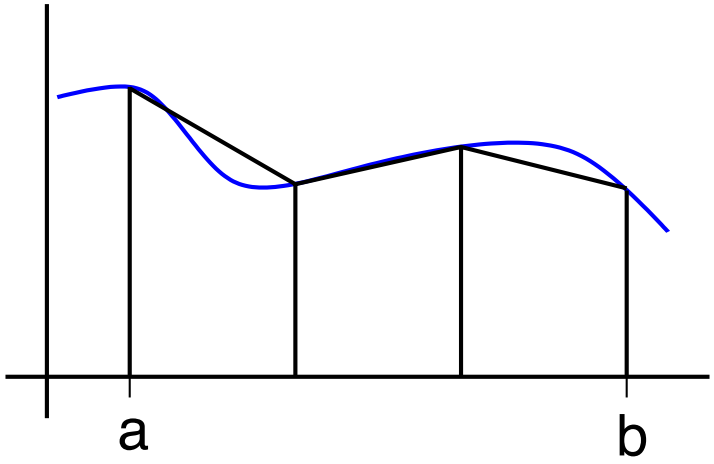
$$\mathbb{E}[V(t)] = \int_{\mathbb{R}^6} V(t,q)\rho(q)dq$$

Numerical Quadrature:

$$\int_{\mathbb{R}^p} f(q)\rho(q)dq \approx \sum_{r=1}^{R} f(q^r)w^r$$

Questions:

- How do we choose the quadrature points and weights?
  - E.g., Newton-Cotes; e.g., trapezoid rule



$$\int_a^b f(q)dq \approx \frac{h}{2}\left[f(a) + f(b) + 2\sum_{r=1}^{R-2} f(q^r)\right]$$

$$q^r = a + hr\ ,\ h = \frac{b-a}{R-1}$$

Numerical Quadrature

Motivation: Computation of expected values requires approximation of integrals

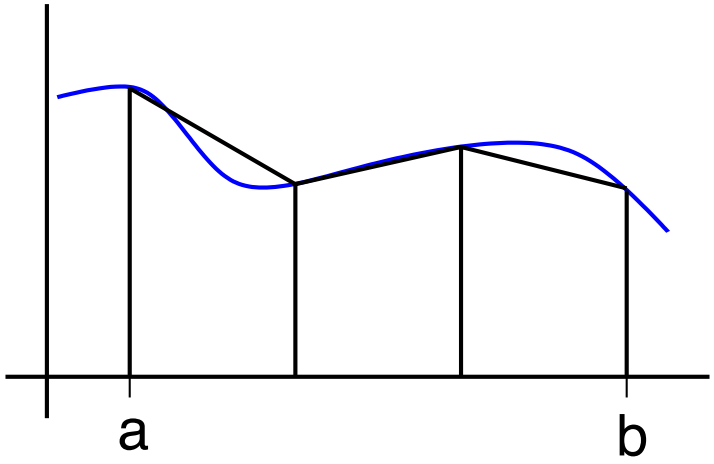
$$\mathbb{E}[u(t,x)] = \int_{\mathbb{R}^p} u(t,x,q)\rho(q)dq$$

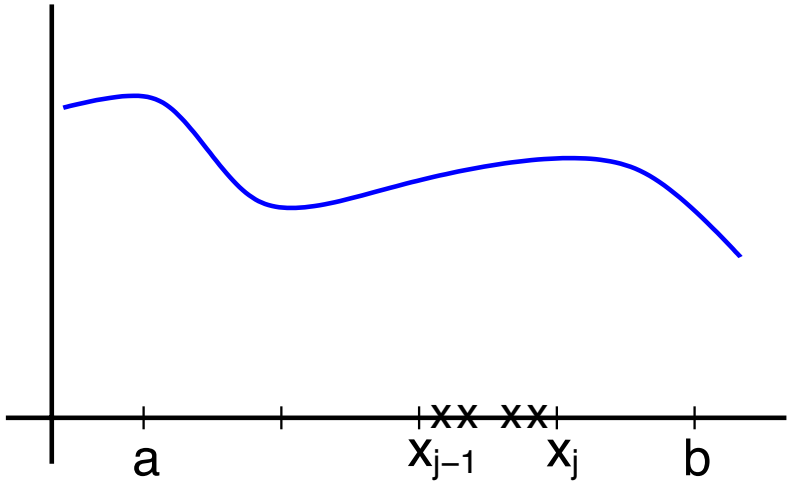
Numerical Quadrature:

$$\int_{\mathbb{R}^p} f(q)\rho(q)dq \approx \sum_{r=1}^{R} f(q^r)w^r$$

Questions:

- How do we choose the quadrature points and weights?
  - E.g., Newton-Cotes, Gaussian algorithms





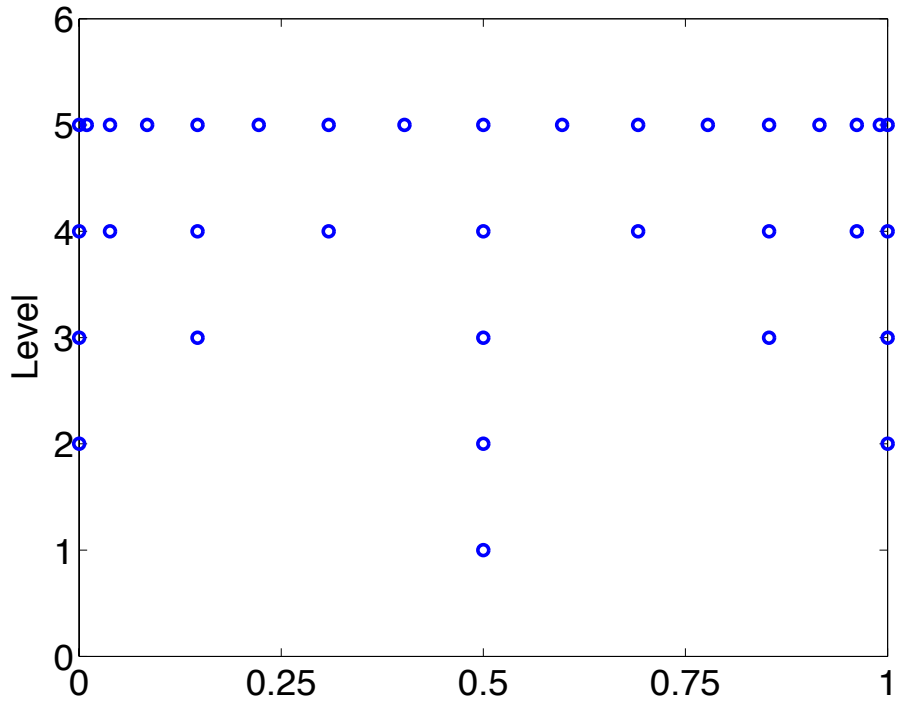
Numerical Quadrature

Numerical Quadrature:

$$\int_{\mathbb{R}^p} f(q)\rho(q)dq \approx \sum_{r=1}^{R} f(q^r)w^r$$

Questions:

- Can we construct nested algorithms to improve efficiency?
  - E.g., employ Clenshaw-Curtis points

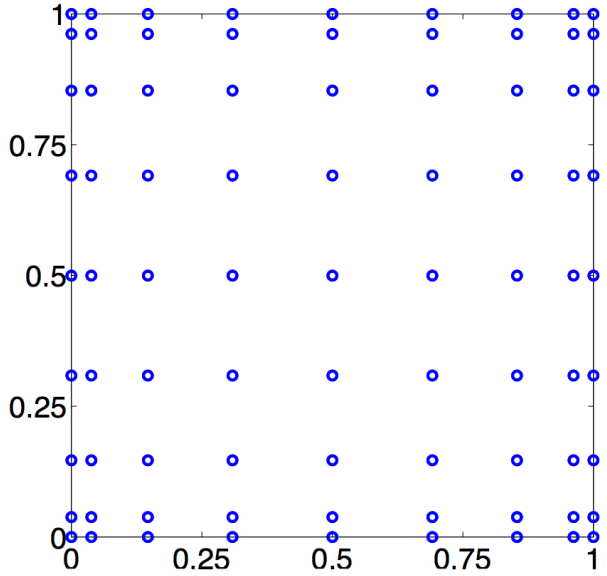


Numerical Quadrature

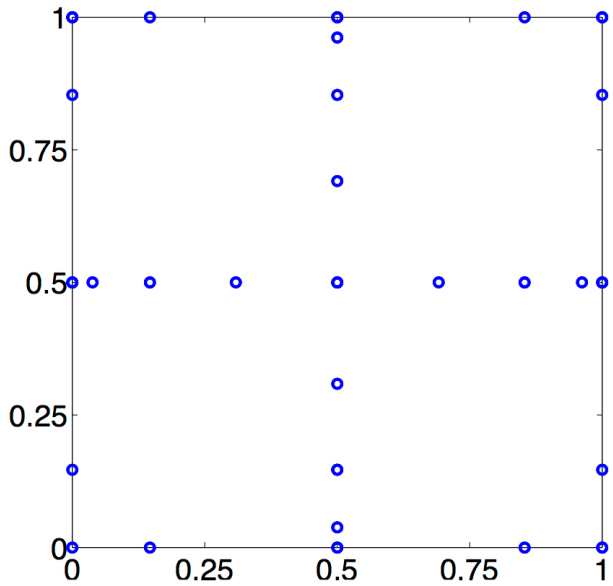
Questions:

- How do we reduce required number of points while maintaining accuracy?

Tensored Grids: Exponential growth



Sparse Grids: Same accuracy



p	R_ℓ	Sparse Grid $\mathcal{R}$	Tensored Grid $R = (R_\ell)^p$
2	9	29	81
5	9	241	59,049
10	9	1581	$> 3 \times 10^9$
50	9	171,901	$> 5 \times 10^{47}$
100	9	1,353,801	$> 2 \times 10^{95}$

Numerical Quadrature

Problem:

- Accuracy of methods diminishes as parameter dimension p increases
- Suppose $f \in C^{\alpha}([0,1]^{p})$
- Tensor products: Take R_{ℓ} points in each dimension so $R = (R_{\ell})^{p}$ total points
- Quadrature errors:

  Newton-Cotes: $E \sim \mathcal{O}(R_{\ell}^{-\alpha}) = \mathcal{O}(R^{-\alpha/p})$

  Gaussian: $E \sim \mathcal{O}(e^{-\beta R_{\ell}}) = \mathcal{O}\left(e^{-\beta \sqrt[p]{R}}\right)$

  Sparse Grid: $E \sim \mathcal{O}\left(\mathcal{R}^{-\alpha} \log(\mathcal{R})^{(p-1)(\alpha+1)}\right)$

Numerical Quadrature

Problem:

- Accuracy of methods diminishes as parameter dimension p increases
- Suppose $f \in C^{\alpha}([0,1]^p)$
- Tensor products: Take R_ℓ points in each dimension so $R = (R_\ell)^p$ total points
- Quadrature errors:

  Newton-Cotes: $E \sim \mathcal{O}(R_\ell^{-\alpha}) = \mathcal{O}(R^{-\alpha/p})$

  Gaussian: $E \sim \mathcal{O}(e^{-\beta R_\ell}) = \mathcal{O}\left(e^{-\beta \sqrt[p]{R}}\right)$

  Sparse Grid: $E \sim \mathcal{O}\left(\mathcal{R}^{-\alpha} \log(\mathcal{R})^{(p-1)(\alpha+1)}\right)$

- Alternative: Monte Carlo quadrature

$$\int_{\mathbb{R}^p} f(q)\rho(q)dq \approx \frac{1}{R}\sum_{r=1}^{R} f(q^r) \quad , \quad E \sim \left(\frac{1}{\sqrt{R}}\right)$$

- Advantage: Errors independent of dimension p
- Disadvantage: Convergence is very slow!

Numerical Quadrature

Problem:

- Accuracy of methods diminishes as parameter dimension p increases
- Suppose $f \in C^{\alpha}([0,1]^p)$
- Tensor products: Take R_ℓ points in each dimension so $R = (R_\ell)^p$ total points
- Quadrature errors:

  Newton-Cotes: $E \sim \mathcal{O}(R_\ell^{-\alpha}) = \mathcal{O}(R^{-\alpha/p})$

  Gaussian: $E \sim \mathcal{O}(e^{-\beta R_\ell}) = \mathcal{O}\left(e^{-\beta \sqrt[p]{R}}\right)$

  Sparse Grid: $E \sim \mathcal{O}\left(\mathcal{R}^{-\alpha} \log(\mathcal{R})^{(p-1)(\alpha+1)}\right)$

- Alternative: Monte Carlo quadrature

$$\int_{\mathbb{R}^p} f(q)\rho(q)dq \approx \frac{1}{R}\sum_{r=1}^{R} f(q^r) \quad , \quad E \sim \left(\frac{1}{\sqrt{R}}\right)$$

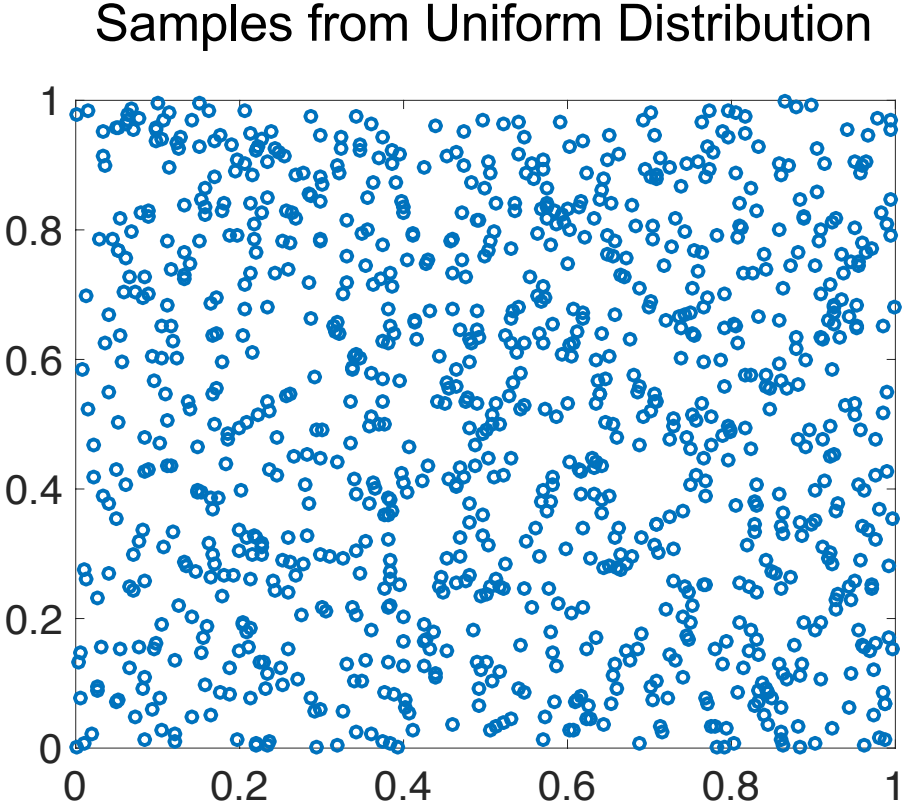
- Advantage: Errors independent of dimension p
- Disadvantage: Convergence is very slow!

Conclusion: For high enough dimension p, monkeys throwing darts will beat Gaussian and sparse grid techniques! Will Cousins, former student of Pierre Gremaud.

Monte Carlo Sampling Techniques

Issues:

- Very low accuracy and slow convergence
- Random sampling may not “randomly” cover space …

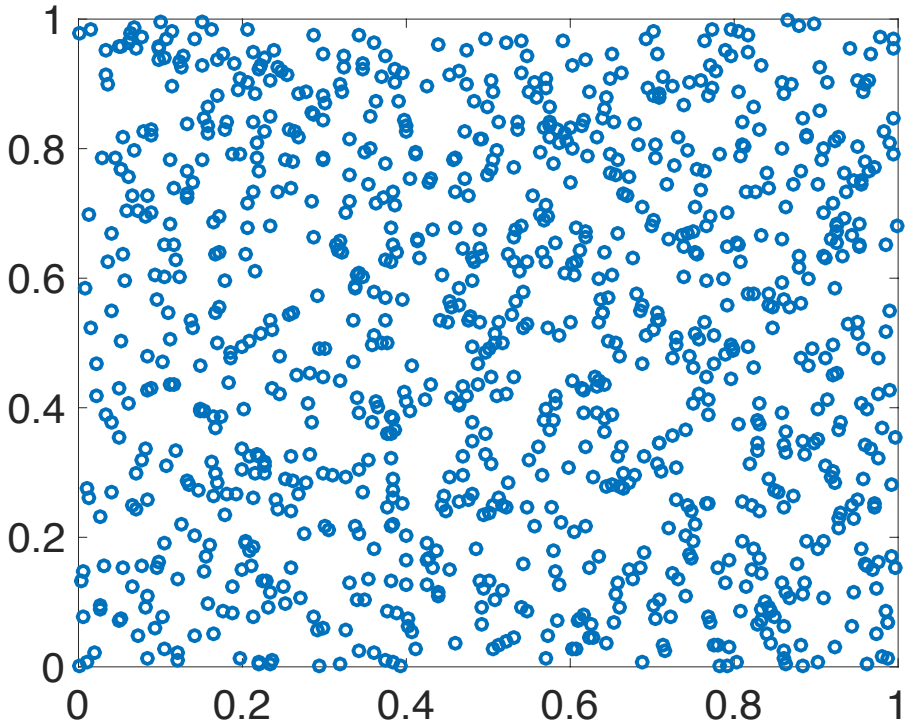


Monte Carlo Sampling Techniques

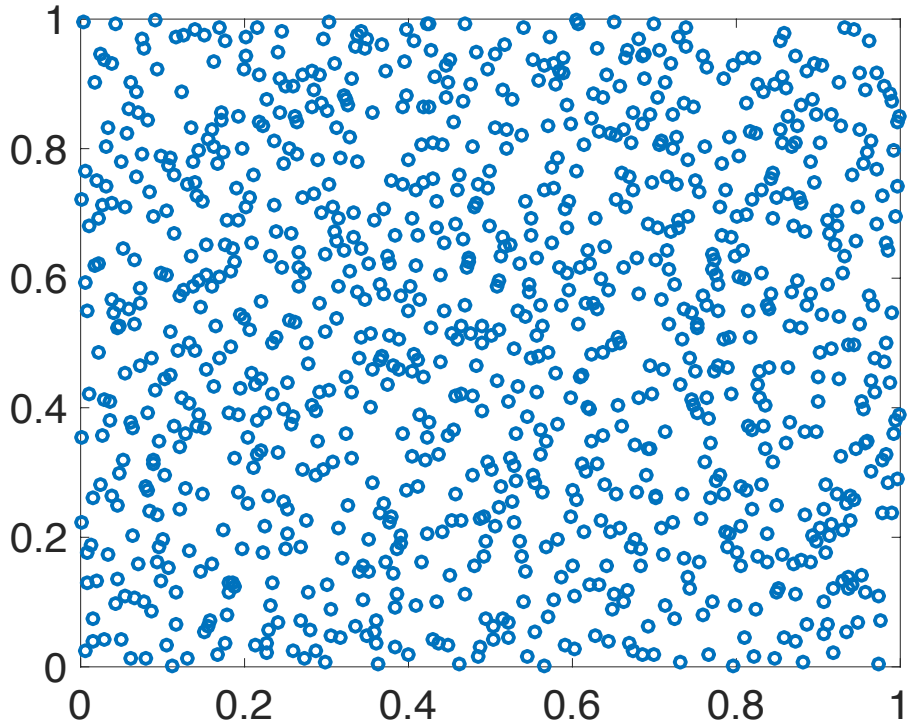
Issues:

- Very low accuracy and slow convergence
- Random sampling may not “randomly” cover space …

Samples from Uniform Distribution



Sobol’ Points



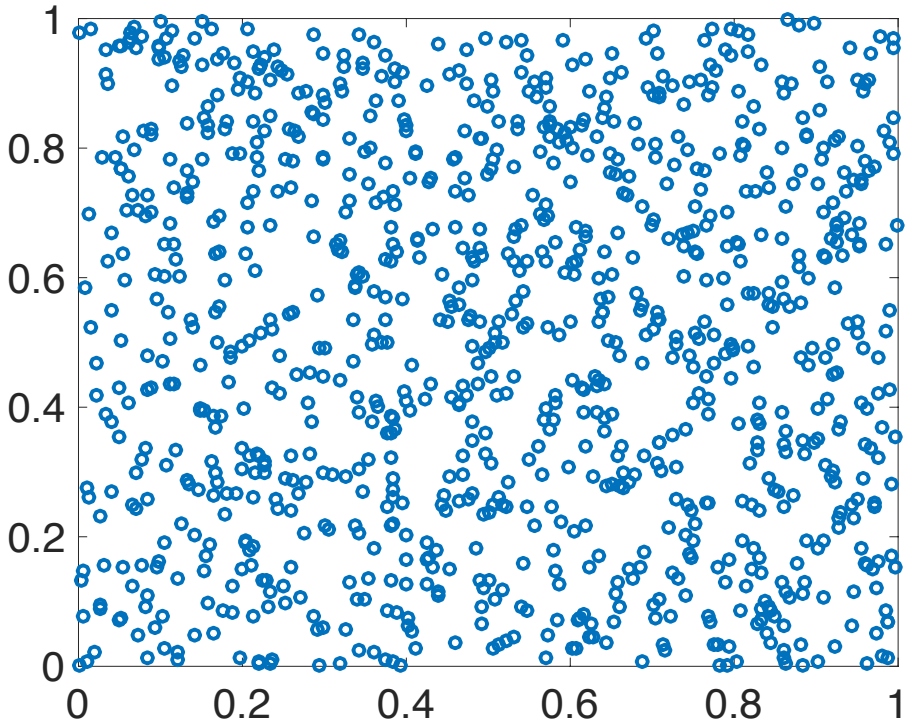
Sobol’ Sequence: Use a base of two to form successively finer uniform partitions of unit interval and reorder coordinates in each dimension.

Quasi-Monte Carlo Sampling Techniques

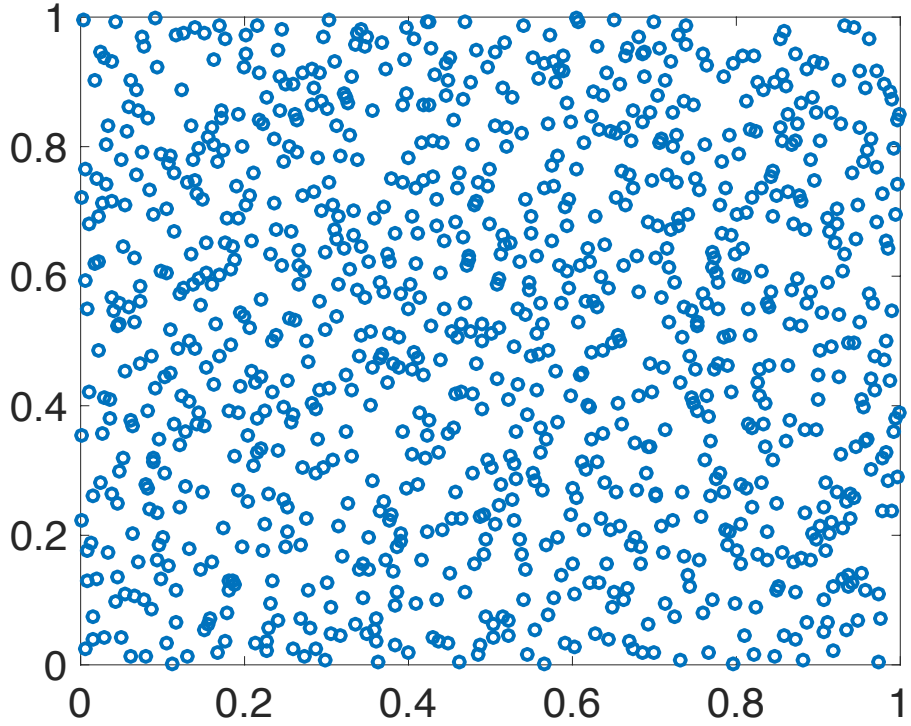
Issues:

- Very low accuracy and slow convergence
- Random sampling may not “randomly” cover space …

Samples from Uniform Distribution



Sobol’ Points



$$\int_{\mathbb{R}^p} f(q)\rho(q)dq \approx \frac{1}{R}\sum_{r=1}^{R} f(q^r) \quad , \quad E \sim \left(\frac{1}{\sqrt{R}}\right)$$

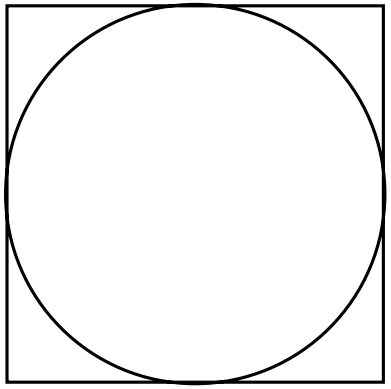
$$E \sim O\left(\frac{(\log R)^p}{R}\right)$$

Monte Carlo Sampling Techniques

Example: Use Monte Carlo sampling to approximate area of circle

$$\frac{A_c}{A_s} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}$$

$$\Rightarrow A_c = \frac{\pi}{4} A_s$$



Strategy:

- Randomly sample N points in square $\Rightarrow$ approximately $N\frac{\pi}{4}$ in circle
- Count M points in circle

  $$\Rightarrow \pi \approx \frac{4M}{N}$$

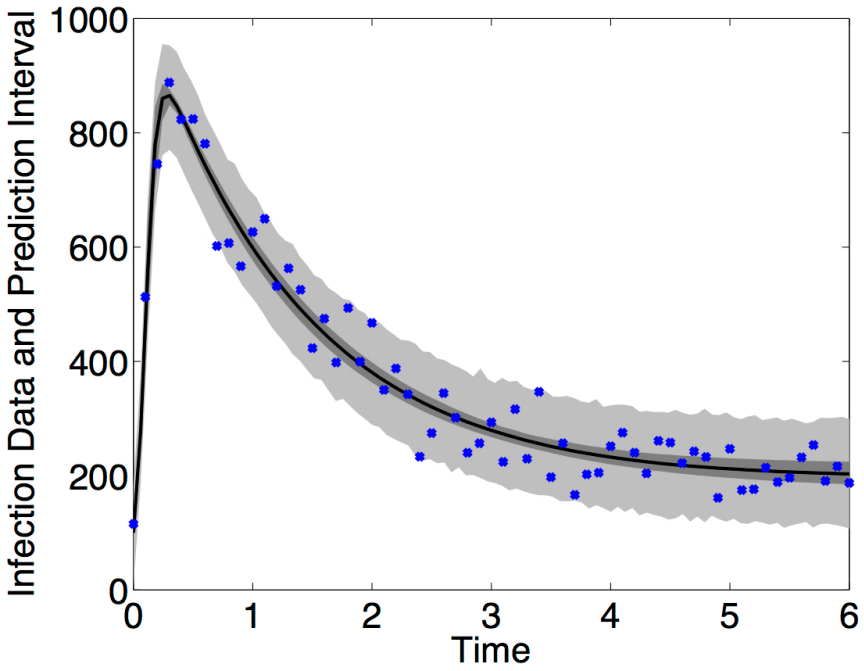
Quasi-Monte Carlo:

- Recent SAMSI Program on *Quasi-Monte Carlo and High Dimensional Sampling Methods in Applied Math* in 2017-18

Confidence, Credible and Prediction Intervals

Note:

- We now know how to compute the mean response for the QoI.
- How do we compute appropriate intervals?



SIR Model:

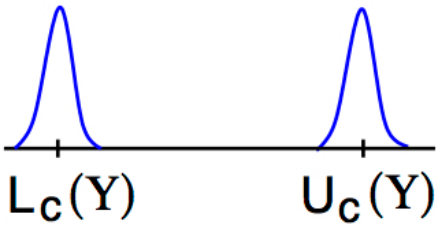
$$\frac{dS}{dt} = \delta N - \delta S - \gamma IS \quad , \; S(0) = S_0$$ Susceptible

$$\frac{dI}{dt} = \gamma IS - (r + \delta)I \quad , \; I(0) = I_0$$ Infectious

$$\frac{dR}{dt} = rI - \delta R \quad , \; R(0) = R_0$$ Recovered

Confidence, Credible and Prediction Intervals

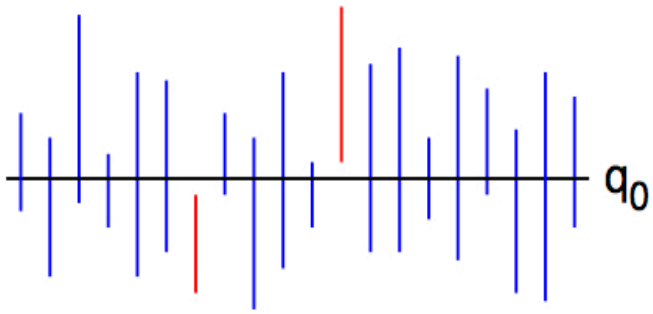
Data: $\Upsilon = [\Upsilon_1, \cdots, \Upsilon_n]$ of iid random observations



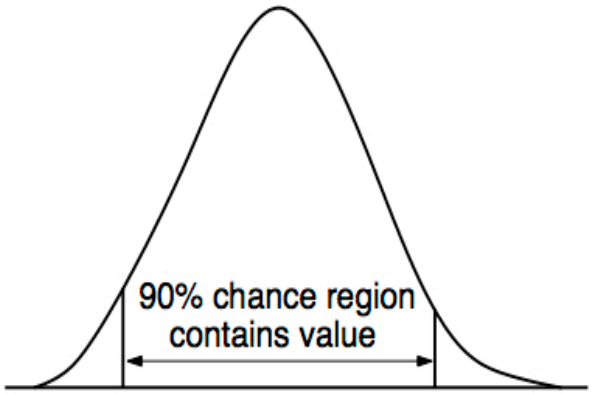
Confidence Interval (Frequentist): A $100 \times (1-\alpha)$% confidence interval for a fixed, unknown parameter q_0 is a random interval $[L_c(\Upsilon), U_c(\Upsilon)]$, having probability at least $1-\alpha$ of covering q_0 under the joint distribution of Υ.

Credible Interval (Bayesian): A $100 \times (1-\alpha)$ % credible interval is that having probability at least $1-\alpha$ of containing q.

Strategy: Sample out of parameter density $\rho_Q(q)$



90% Confidence Intervals



90% Credible Interval

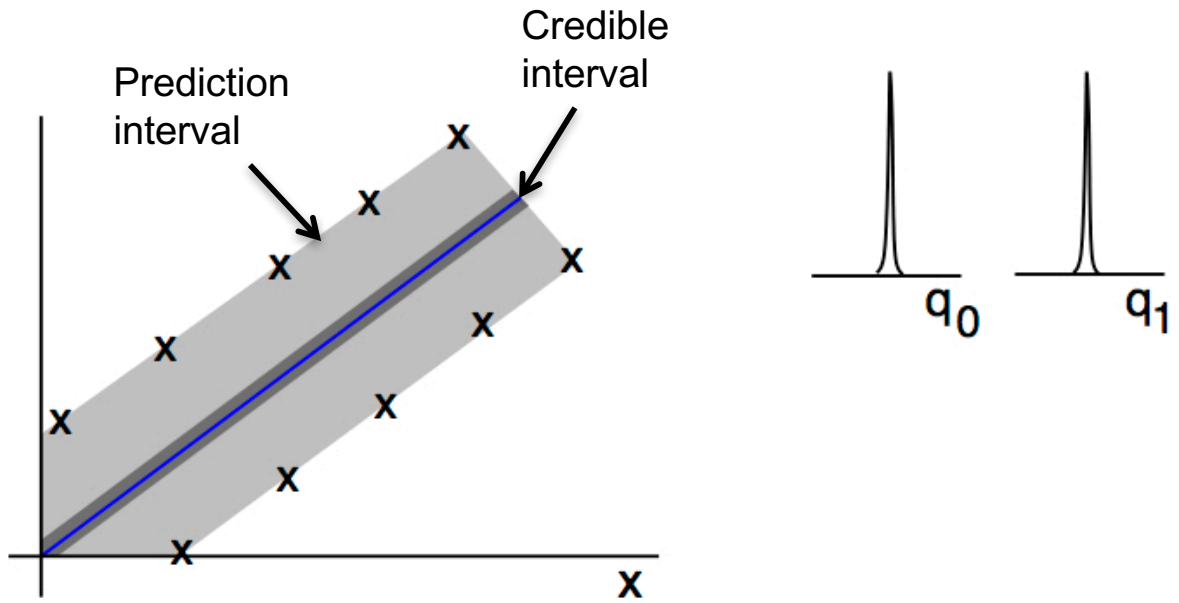
Confidence, Credible and Prediction Intervals

Data: $\Upsilon = [\Upsilon_1, \cdots, \Upsilon_n]$ of iid random observations

Prediction Interval: A $100 \times (1-\alpha)$% prediction interval for a future observable Υ_{n+1} is a random interval $[L_c(\Upsilon), U_c(\Upsilon)]$ having probability at least $1-\alpha$ of of containing Υ_{n+1} under the joint distribution of $(\Upsilon, \Upsilon_{n+1})$.

Example: Consider linear model

$$\Upsilon_i = q_0 + q_1 x_i + \varepsilon_i \ , \ i = 1, \cdots, n$$



Example: HIV Model

Model:

$$\dot{T}_1 = \lambda_1 - d_1 T_1 - (1-\varepsilon) k_1 V T_1$$

$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1-f\varepsilon) k_2 V T_2$$

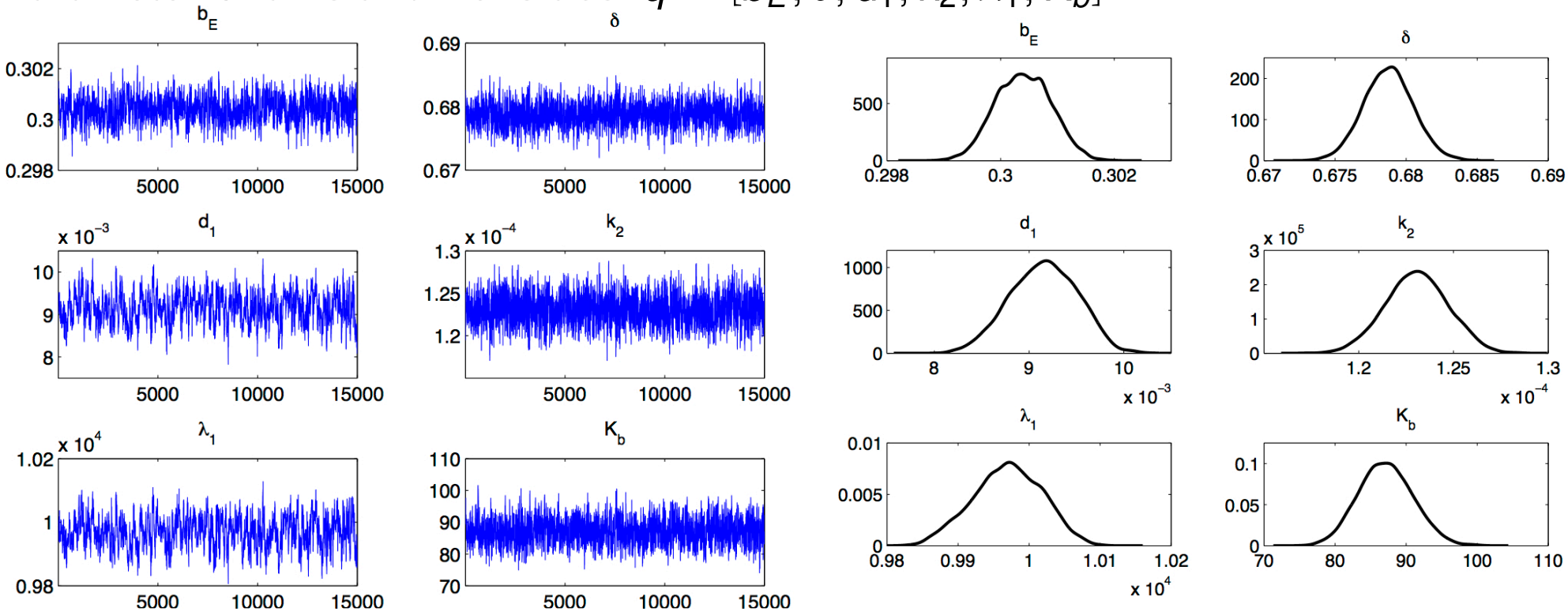
$$\dot{T}_1^* = (1-\varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1-f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

$$\dot{V} = N_T \delta (T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1-f\varepsilon)\rho_2 k_2 T_2] V$$

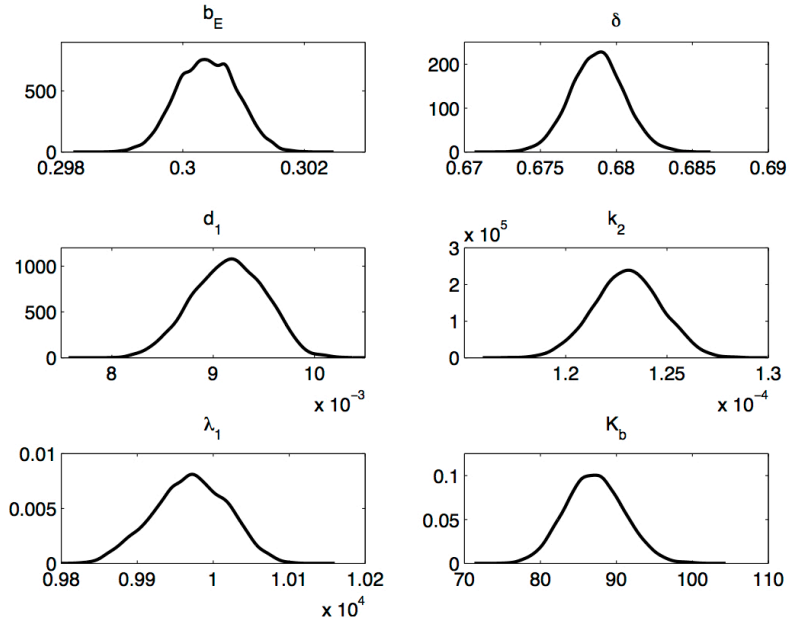
$$\dot{E} = \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E$$

Parameter Chains and Densities: $q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$



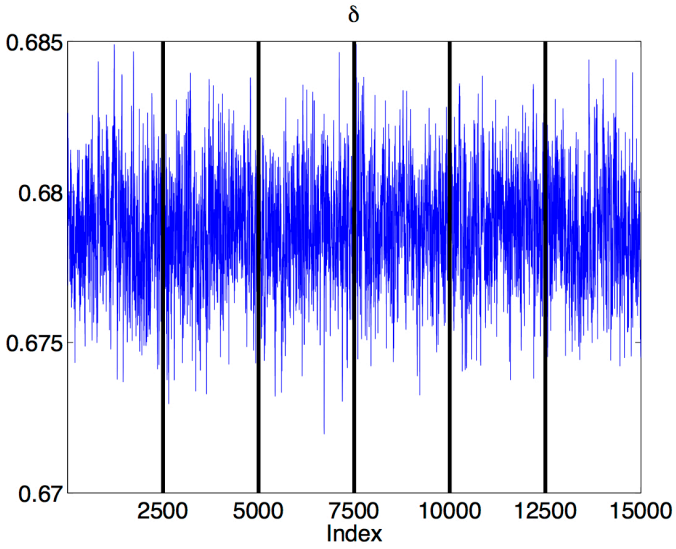
Propagation of Uncertainty – HIV Example

Parameter Densities:

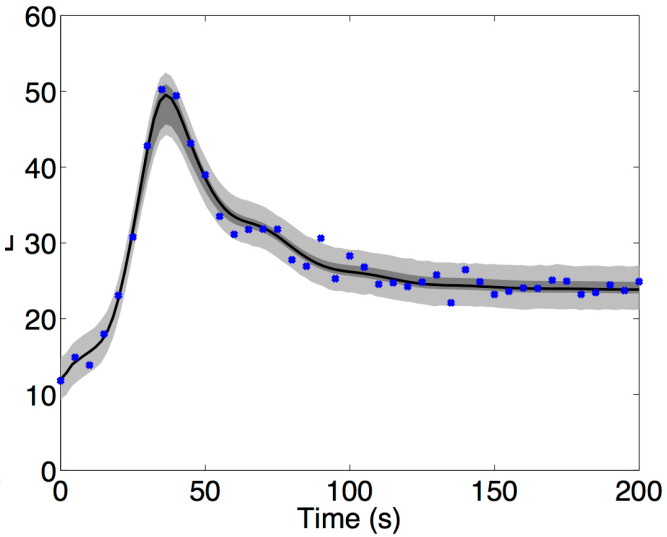


Techniques:

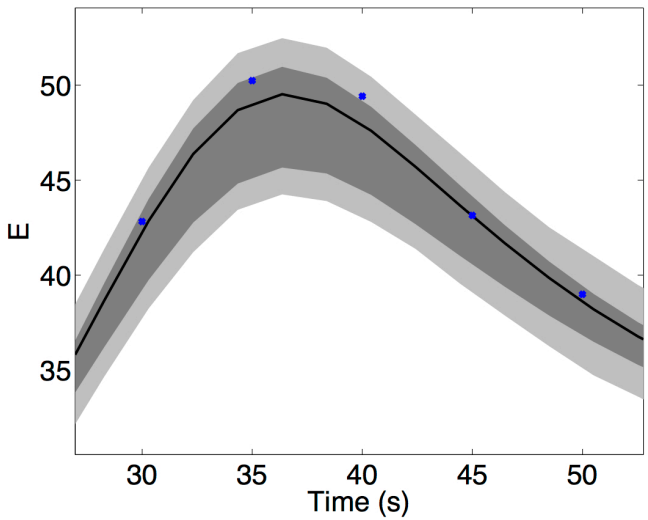
- Sample from parameter and observation error densities to construct mean response, credible intervals, and prediction intervals for QoI.
- Slow convergence rate $\mathcal{O}(1/\sqrt{M})$



Samples from Chain



Data, Credible Intervals and Prediction Intervals



Non-Gaussian Credible and Prediction Intervals

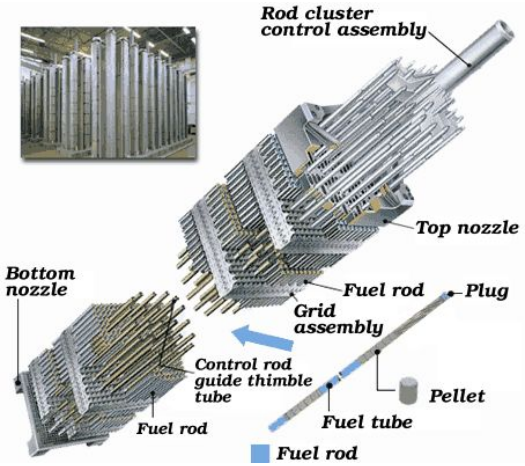
Use of Prediction Intervals: Nuclear Power Plant Design

Subchannel Code (COBRA-TF): numerous closure relations, ~70 parameters

e.g., Dittus—Boelter Relation

$$Nu = 0.023 Re^{0.8} Pr^{0.4}$$

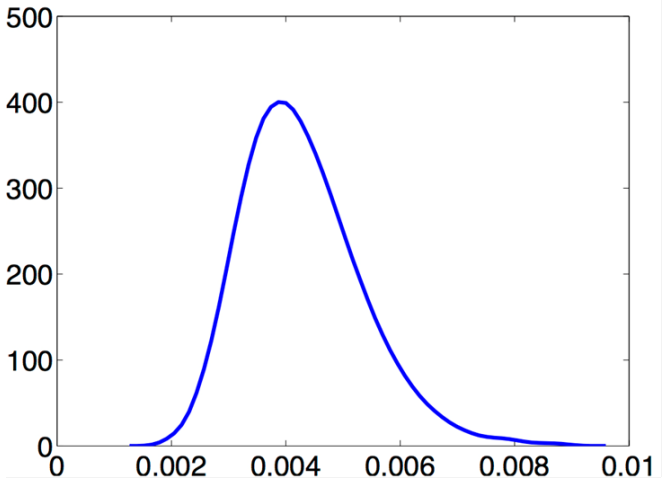
Nu: Nusselt number
Re: Reynolds number
Pr: Prandtl number



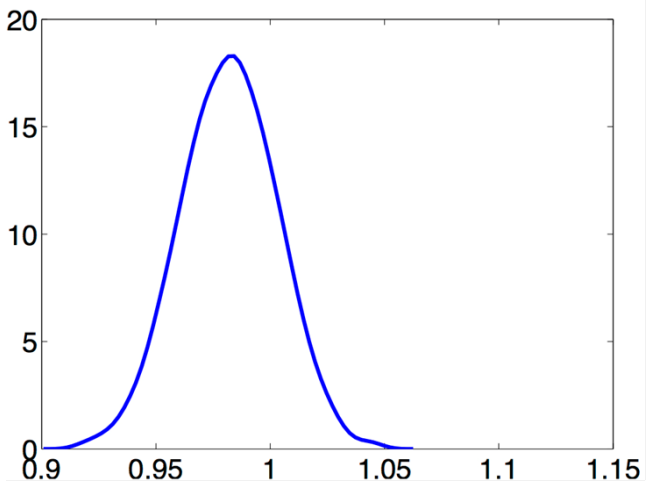
Industry Standard: Employ conservative, uniform, bounds

i.e., [0, 0.046], [0, 1.6], [0,0.8]

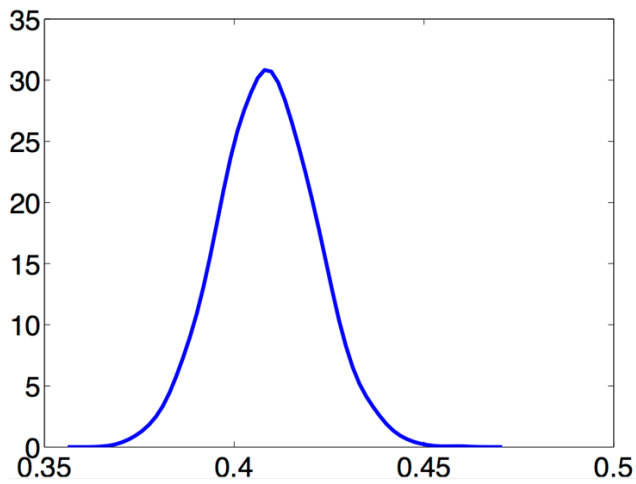
Bayesian Analysis: Employ conservative bounds as priors



$2\sigma \approx 0.0035$



$2\sigma \approx 0.06$

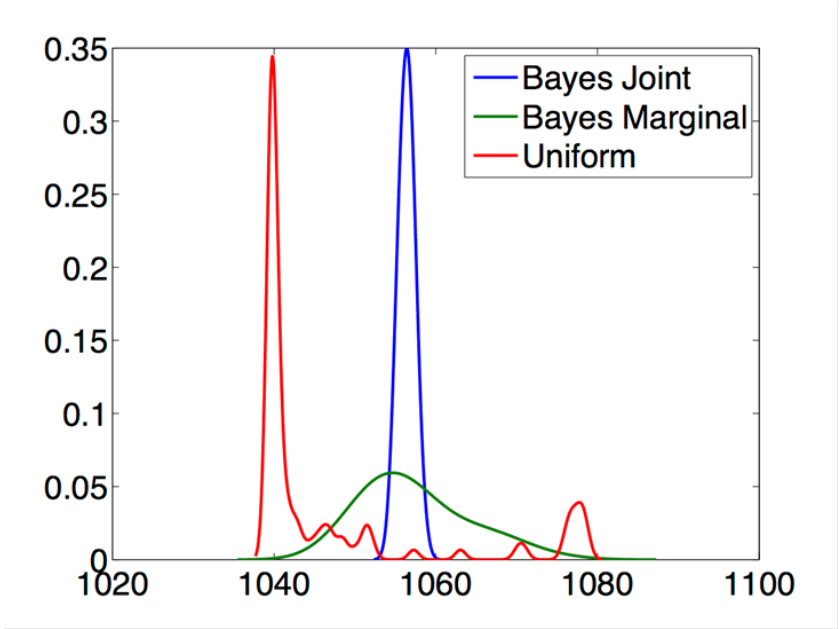


$2\sigma \approx 0.03$

Note: Substantial reduction in parameter uncertainty

Use of Prediction Intervals: Nuclear Power Plant Design

Strategy: Propagate parameter uncertainties through COBRA-TF to determine uncertainty in maximum fuel temperature



Notes:

- Temperature uncertainty reduced from 40 degrees to 5 degrees
- Can run plant 20 degrees hotter, which significantly improves efficiency

Ramification: Savings of 10 billion dollars per year for US power plants

Issues:

- We considered only one of many closure relations
- Nuclear regulatory commission takes years to change requirements and codes

Good News: We are now working with Westinghouse to reduce uncertainties.

Practicum Problem

SIR Model:

$$\frac{dS}{dt} = \delta N - \delta S - \gamma k I S \quad , \; S(0) = S_0$$

$$\frac{dI}{dt} = \gamma k I S - (r + \delta) I \quad , \; I(0) = I_0$$

$$\frac{dR}{dt} = rI - \delta R \quad , \; R(0) = R_0$$

Note:

- Run either the 3 or 4 parameter model and compute the prediction intervals.

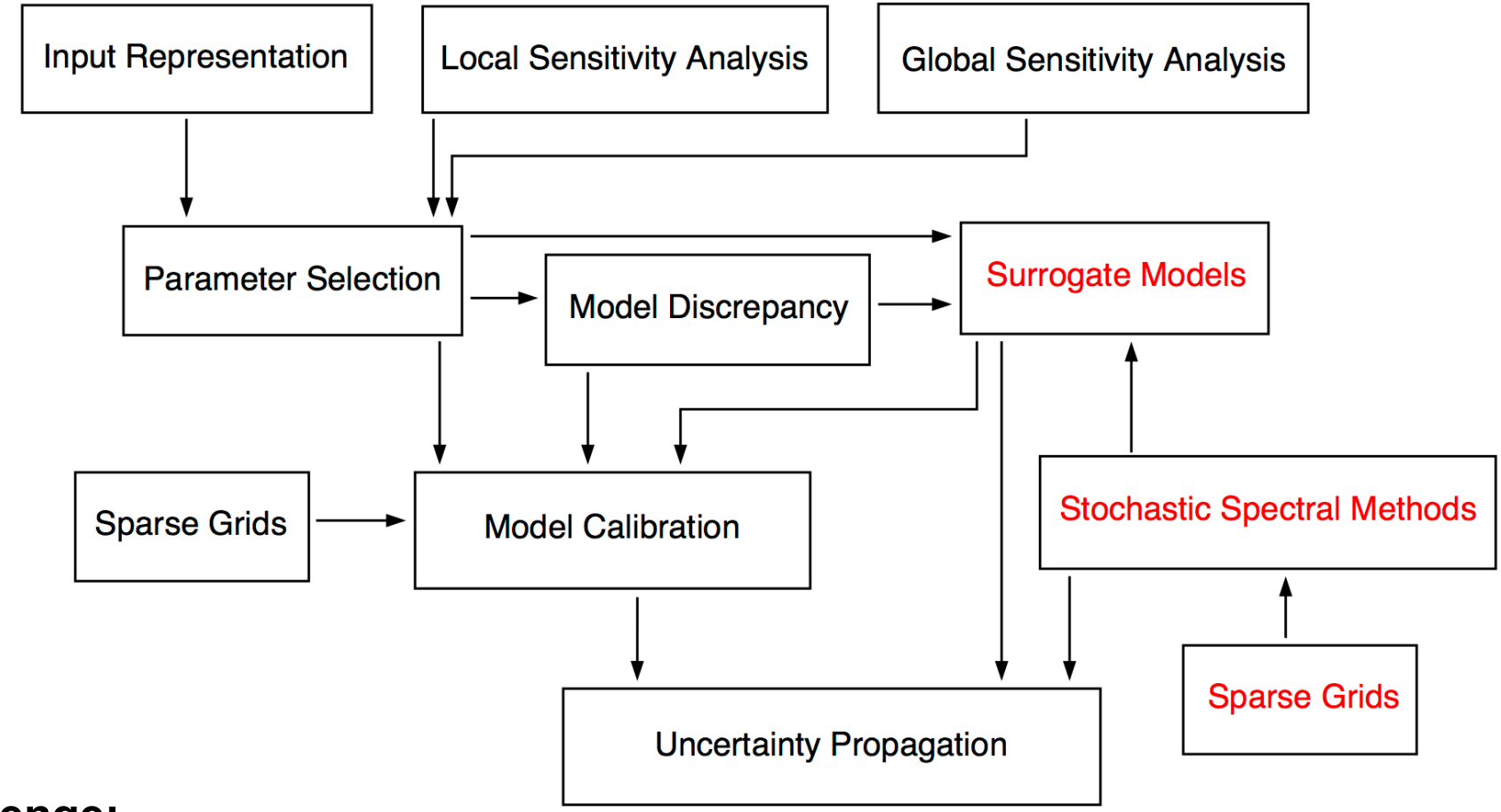
Monte Carlo Quadrature:

- Run rand_points.m to observe uniformly sampled and Sobol' points.
- Run pi_approx.m with different values of N to see if you observe convergence rate of $1/\sqrt{N}$

Website:

- http://www4.ncsu.edu/~rsmith/RTG_BIOMATH18/

Steps in Uncertainty Quantification



Challenge:

- How do we do uncertainty quantification for computationally expensive models?
- Example:
  - We have a computational budget of 5000 model evaluations.
  - Bayesian inference and uncertainty propagation require 120,000 evaluations.

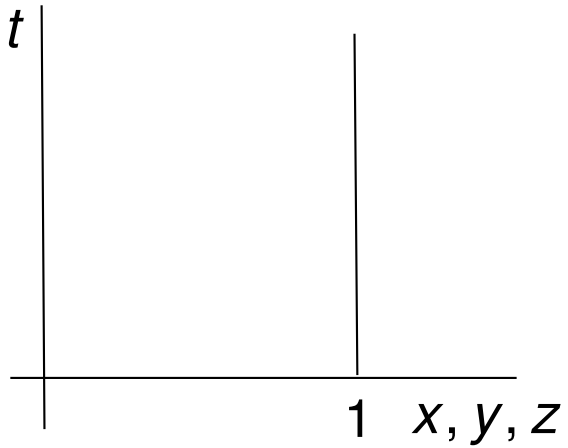
Surrogate Models: Motivation

Example: Consider the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + f(q)$$

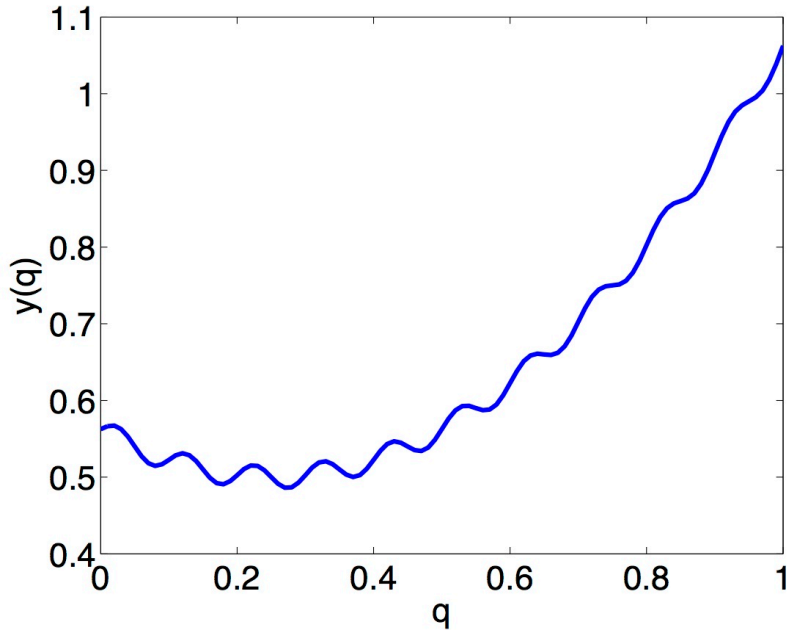
Boundary Conditions

Initial Conditions



with the response

$$y(q) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 u(t, x, y, z)\, dx dy dz dt$$



Notes:

- Requires approximation of PDE in 3-D
- What would be a simple surrogate?

Surrogate Models: Motivation

Example: Consider the heat equation

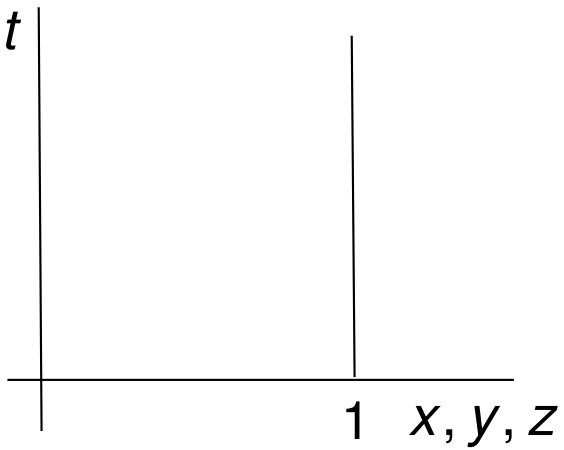
$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + f(q)$$

Boundary Conditions

Initial Conditions

with the response

$$y(q) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 u(t, x, y, z)\, dx dy dz dt$$

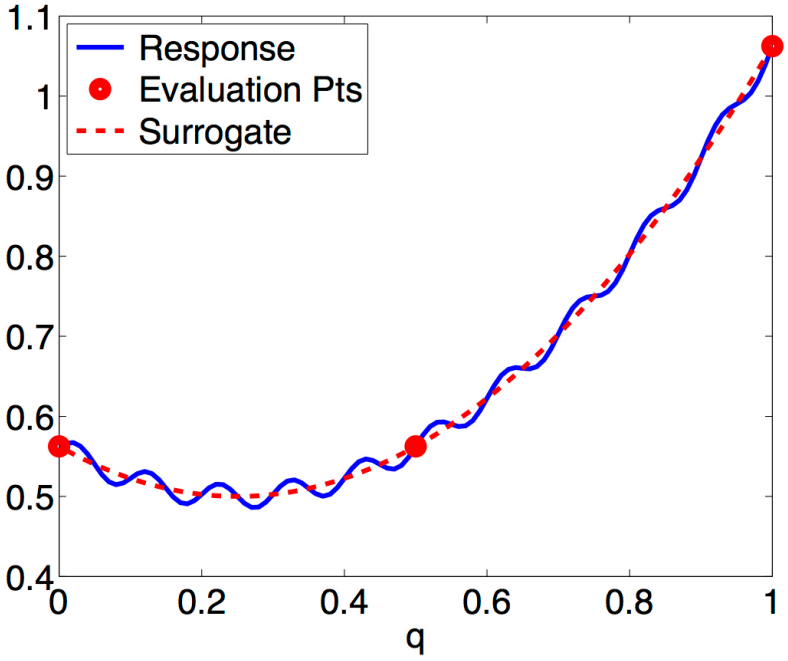


Surrogate: Quadratic

$$y_s(q) = (q - 0.25)^2 + 0.5$$

Question: How do you construct a polynomial surrogate?

- Regression
- Interpolation



Surrogate Models

Recall: Consider the model

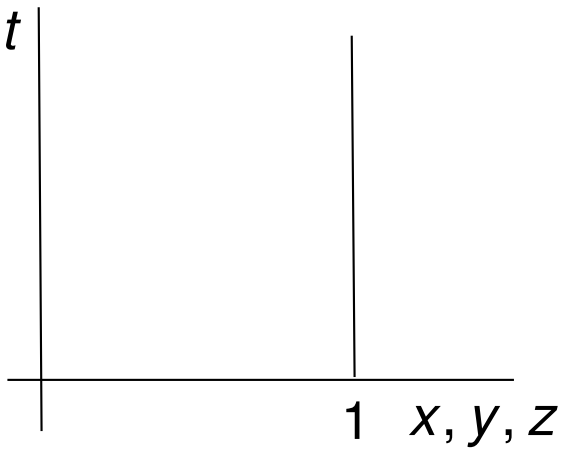
$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + f(q)$$

Boundary Conditions

Initial Conditions

with the response

$$y(q) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 u(t, x, y, z)\, dx dy dz dt$$

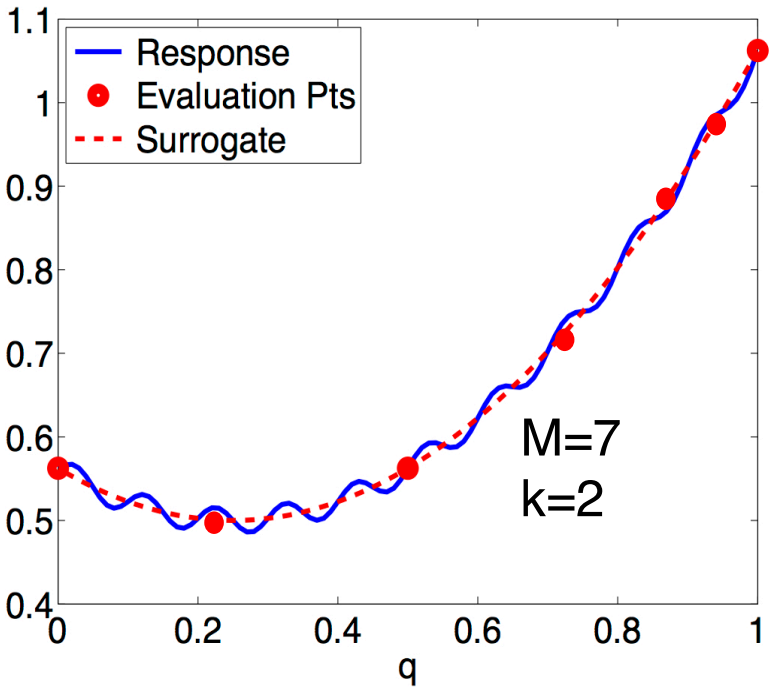


Surrogate: Quadratic

$$y_s(q) = (q - 0.25)^2 + 0.5$$

Question: How do you construct a polynomial surrogate?

- Interpolation
- Regression



Data-Fit Models

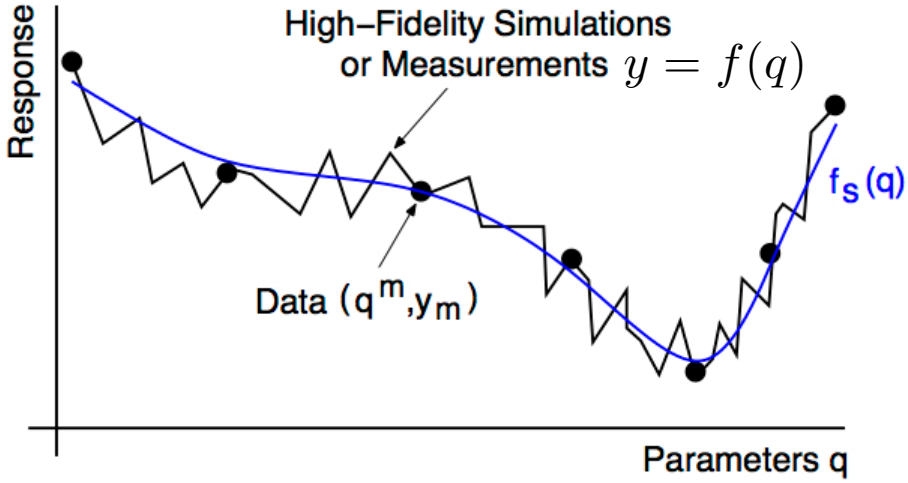
Notes:

• Often termed response surface models, surrogates, emulators, meta-models.

• Rely on interpolation or regression.

• Data can consist of high-fidelity simulations or experiments.

• Common techniques: polynomial models, Gaussian process (Dirk Husmeier), orthogonal polynomials.



Strategy: Consider high fidelity model

$$y = f(q)$$

with M model evaluations

$$y_m = f(q^m)\ ,\ \ m = 1, \ldots, M$$

Statistical Model: $f_s(q)$: Surrogate for $f(q)$

$$y_m = f_s(q^m) + \varepsilon_m\ ,\ \ m = 1, \ldots, M$$

Surrogate:

$$y^K(Q) = f_s(Q) = \sum_{k=0}^{K} \alpha_k \Psi_k(Q)$$

Note: $\Psi_k(Q)$ orthogonal with respect to inner product associated with pdf

e.g., $Q \sim N(0,1)$: Hermite polynomials

$Q \sim U(-1,1)$: Legendre polynomials

Orthogonal Polynomial Representations

Representation:

$$y^K(Q) = \sum_{k=0}^{K} \alpha_k \Psi_k(Q)$$

Note: $\Psi_0(Q) = 1$ implies that

$$\mathbb{E}[\Psi_0(Q)] = 1$$

$$\mathbb{E}[\Psi_i(Q)\Psi_j(Q)] = \int_\Gamma \Psi_i(q)\Psi_j(q)\rho(q)dq$$

$$= \delta_{ij}\gamma_i$$

where $\gamma_i = \mathbb{E}[\Psi_i^2(Q)]$

Properties:

(i) $\mathbb{E}[y^K(Q)] = \alpha_0$

(ii) $\text{var}[y^K(Q)] = \sum_{k=1}^{K} \alpha_k^2 \gamma_k$

Note: Can be used for

- Uncertainty propagation
- Sobol-based global sensitivity analysis

Issue: How does one compute α_k , $k = 0, \ldots, K$?

- Stochastic Galerkin techniques (Polynomial Chaos Expansion – PCE)
- Nonintrusive PCE (Discrete projection)
- Stochastic collocation
- Regression-based methods with sparsity control (Lasso)

Note: Methods nonintrusive and treat code as blackbox.

Orthogonal Polynomial Representations:

Nonintrusive PCE: Take weighted inner product of $y(q) = \sum_{k=0}^{\infty} \alpha_k \Psi_k(q)$ to obtain

$$\alpha_k = \frac{1}{\gamma_k} \int_\Gamma y(q) \Psi_k(q) \rho(q) dq$$

Quadrature:

$$\alpha_k \approx \frac{1}{\gamma_k} \sum_{r=1}^{R} y(q^r) \Psi_k(q^r) w^r$$

Note:

(i) Low-dimensional: Tensored 1-D quadrature rules – e.g., Gaussian

(ii) Moderate-dimensional: Sparse grid (Smolyak) techniques

(iii) High-dimensional: Monte Carlo or quasi-Monte Carlo (QMC) techniques

Regression-Based Methods with Sparsity Control (Lasso): Solve

$$\min_{\alpha \in \mathbb{R}^{K+1}} \|\Lambda \alpha - d\|^2 \quad \text{subject to} \quad \sum_{k=0}^{K} |\alpha_k| \leqslant \tau$$

Note: Sample points $\{q^m\}_{m=1}^{M}$

$\Lambda \in \mathbb{R}^{M \times (K+1)}$ where $\Lambda_{jk} = \Psi_k(q^j)$

$d = [y(q^1), \ldots, y(q^m)]$

e.g., SPGL1

• MATLAB Solver for large-scale sparse reconstruction

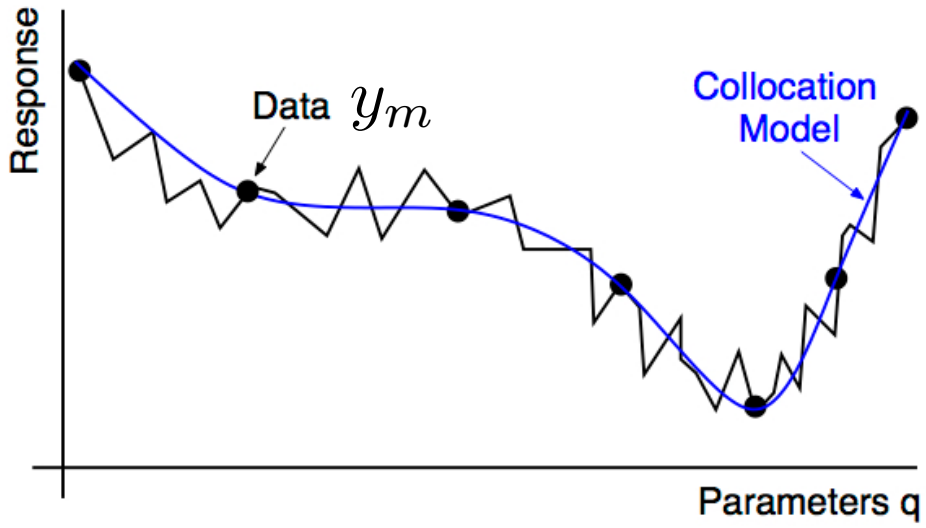
Stochastic Collocation

Strategy: Consider high fidelity model

$$y = f(q)$$

with M model evaluations

$$y_m = f(q^m)\ ,\ m = 1, \ldots, M$$



Collocation Surrogate:

$$Y^M(q) = \sum_{m=1}^{M} y_m L_m(q)$$

where $L_m(q)$ is a Lagrange polynomial, which in 1-D, is represented by

$$L_m(q) = \prod_{\substack{j=0 \\ j \neq m}}^{M} \frac{q - q^j}{q^m - q^j} = \frac{(q-q^1)\cdots(q-q^{m-1})(q-q^{m+1})\cdots(q-q^M)}{(q^m-q^1)\cdots(q^m-q^{m-1})(q^m-q^{m+1})\cdots(q^m-q^M)}$$

Note:

$$L_m(q^j) = \delta_{jm} = \begin{cases} 0 & , \quad j \neq m \\ 1 & , \quad j = m \end{cases}$$

Result: $Y^M(q^m) = y_m$

Example: SIR Cholera Model

Model:

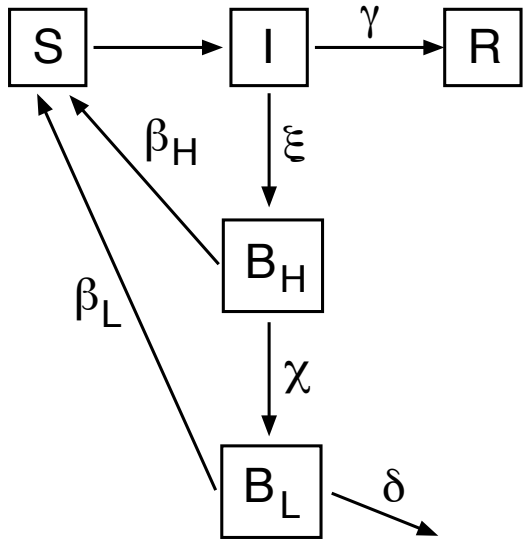
$$\frac{dS}{dt} = bN - \beta_L S\frac{B_L}{\kappa_L + B_L} - \beta_H S\frac{B_H}{\kappa_H + B_H} - bS$$

$$\frac{dI}{dt} = \beta_L S\frac{B_L}{\kappa_L + B_L} + \beta_H S\frac{B_H}{\kappa_H + B_H} - (\gamma + b)I$$

$$\frac{dR}{dt} = \gamma I - bR$$

$$\frac{dB_H}{dt} = \xi I - \chi B_H$$

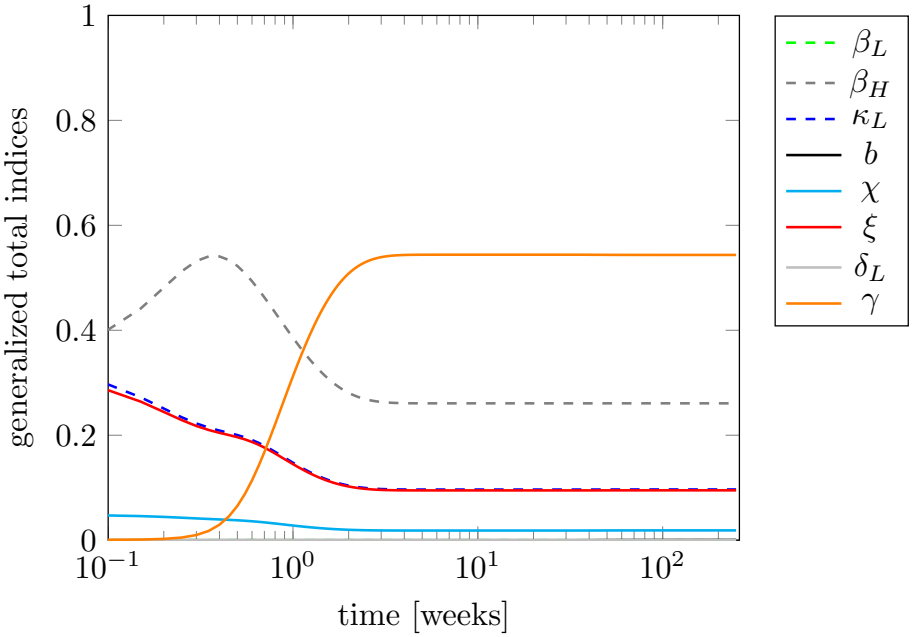
$$\frac{dB_L}{dt} = \chi B_H - \delta B_L$$



Model Parameter	Symbol	Units	Values
Rate of drinking B_L cholera	β_L	$\frac{1}{\text{week}}$	1.5
Rate of drinking B_H cholera	β_H	$\frac{1}{\text{week}}$	7.5 $(*)$
B_L cholera carrying capacity	κ_L	$\frac{\#\text{ bacteria}}{m\ell}$	10^6
B_H cholera carrying capacity	κ_H	$\frac{\#\text{ bacteria}}{m\ell}$	$\frac{\kappa_L}{700}$
Human birth and death rate	b	$\frac{1}{\text{week}}$	$\frac{1}{1560}$
Rate of decay from B_H to B_L	χ	$\frac{1}{\text{week}}$	$\frac{168}{5}$
Rate at which infectious individuals spread B_H bacteria to water	ξ	$\frac{\#\text{ bacteria}}{\#\text{ individuals}\cdot m\ell\cdot\text{week}}$	70
Death rate of B_L cholera	δ	$\frac{1}{\text{week}}$	$\frac{7}{30}$
Rate of recovery from cholera	γ	$\frac{1}{\text{week}}$	$\frac{7}{5}$

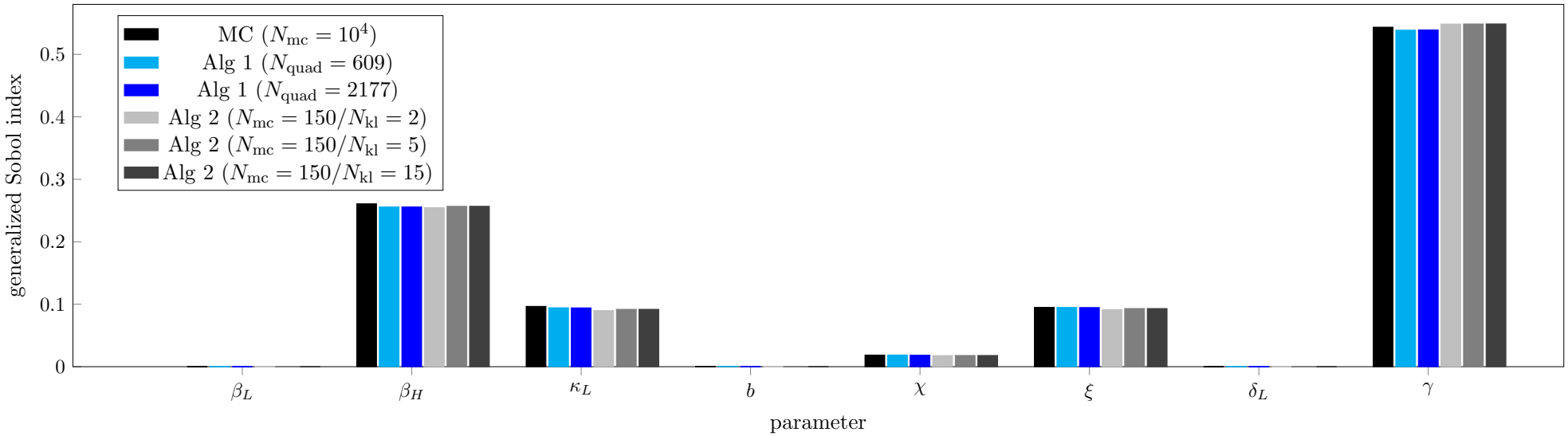
Example: SIR Cholera Model

Strategy: Employed collocation and discrete projection with sparse grids to compute time-dependent global sensitivity indices (Alexanderian and Gremaud)



Conclusion: Sensitive indices

- γ: Recovery rate
- β_H: Rate of drinking B_H cholera
- κ_L: B_L carrying capacity; Note $\kappa_H = \kappa_L/700$
- ξ: Rate at which B_H bacteria spread



Example: SIR Cholera Model

Model:

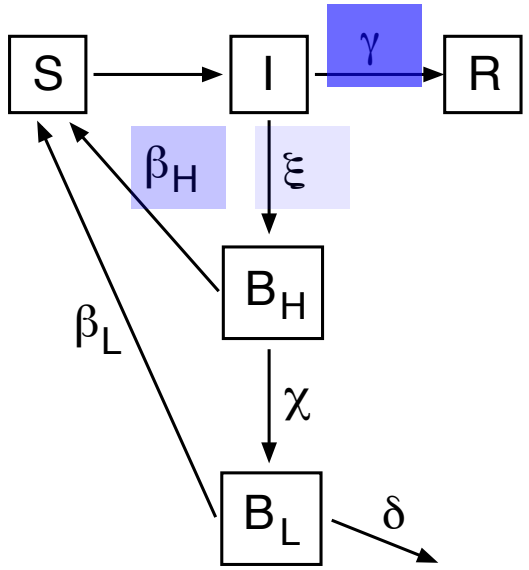
$$\frac{dS}{dt} = bN - \beta_L S\frac{B_L}{\kappa_L + B_L} - \beta_H S\frac{B_H}{\kappa_H + B_H} - bS$$

$$\frac{dI}{dt} = \beta_L S\frac{B_L}{\kappa_L + B_L} + \beta_H S\frac{B_H}{\kappa_H + B_H} - (\gamma + b)I$$

$$\frac{dR}{dt} = \gamma I - bR$$

$$\frac{dB_H}{dt} = \xi I - \chi B_H$$

$$\frac{dB_L}{dt} = \chi B_H - \delta B_L$$



Model Parameter	Symbol	Units	Values
Rate of drinking B_L cholera	β_L	$\frac{1}{\text{week}}$	1.5
Rate of drinking B_H cholera	β_H	$\frac{1}{\text{week}}$	7.5 $(*)$
B_L cholera carrying capacity	κ_L	$\frac{\#\text{ bacteria}}{m\ell}$	10^6
B_H cholera carrying capacity	κ_H	$\frac{\#\text{ bacteria}}{m\ell}$	$\frac{\kappa_L}{700}$
Human birth and death rate	b	$\frac{1}{\text{week}}$	$\frac{1}{1560}$
Rate of decay from B_H to B_L	χ	$\frac{1}{\text{week}}$	$\frac{168}{5}$
Rate at which infectious individuals spread B_H bacteria to water	ξ	$\frac{\#\text{ bacteria}}{\#\text{ individuals}\cdot m\ell\cdot\text{week}}$	70
Death rate of B_L cholera	δ	$\frac{1}{\text{week}}$	$\frac{7}{30}$
Rate of recovery from cholera	γ	$\frac{1}{\text{week}}$	$\frac{7}{5}$

Concluding Remarks

Notes:

- UQ requires a synergy between engineering, statistics, and applied mathematics.
- Model calibration, model selection, uncertainty propagation and experimental design are natural in a Bayesian framework.
- Goal is to predict model responses with quantified and reduced uncertainties.
- Parameter selection is critical to isolate identifiable and influential parameters.
- Surrogate models critical for computationally intensive simulation codes.
- Codes and packages: Sandia Dakota, R, MATLAB, Python, nanoHUB.
- *Prediction is very difficult, especially if it's about the future*, Niels Bohr.