Uncertainty Analysis for Complex Systems: Algorithms and Data -- II

Dongbin Xiu

Department of Mathematics, Purdue University

PURDUE
UNIVERSITY

Support: AFOSR FA9550-08-1-0353
NSF CAREER DMS-0645035
NSF IIS-091447
DOE/NNSA DE-FC52-08NA28617

Overview

- **Uncertainty quantification and stochastic modeling**

- **Algorithms: Uncertainty propagation**
  - Generalized polynomial chaos (gPC)
  - Aleatory and <u>epistemic uncertainty</u>

- **Data: Any data can be helpful**
  - <u>Parameter estimation</u>: inverse inference
  - Uncertainty in models: data assimilation
  - gPC based "free" algorithms

- **Key Issues:**
  - **Efficiency**
  - **Curse of dimensionality**

Uncertainty Propagation: Setup

- **Stochastic PDE:**

$$\frac{\partial u}{\partial t}(t,x,Z) = \mathcal{L}(u) \quad + \text{ boundary/initial conditions}$$

- Uncertain inputs are characterized by n_z random variables Z

- **Input parameterization:**
  - Random parameters
  - Random processes/fields: KL expansion etc.

- ***Aleatory*** **uncertainty:**
  - Distribution of Z is known $F_Z(s) = \Pr(Z \le s), \quad s \in \mathbb{R}^{n_z}$
  - Independence requirement (can be relaxed)

- ***Epistemic*** **uncertainty:**
  - Distribution of Z is not fully known
  - Incomplete PDF, correlation, dependence, etc.
  - *Jakeman, Eldred , Xiu, J. Comput. Phys., vol. 229, 2010*

Epistemic Uncertainty: Setup

- **Governing Equation:**

$$\begin{cases} \dfrac{\partial v}{\partial t}(t,x,Z) = \mathcal{L}(v), & D \times (0,T] \times I_Z \\ \mathcal{B}(v) = 0, & \partial D \times [0,T] \times I_Z \\ v = v_0, & D \times \{t=0\} \times I_Z \end{cases}$$

$$v(x,t,Z) : \bar{D} \times [0,T] \times I_Z \to \mathbb{R} \qquad I_Z \subseteq \mathbb{R}^d$$

- **Epistemic uncertainty:**
    - Distribution of Z is not fully known

- **“Some” prior knowledge:**

$$I_{Z_i} = [\alpha_i, \beta_i], \quad -\infty \le \alpha_i < \beta_i \le \infty$$

$$I_Z \subseteq \mathop{\times}_{i=1}^{d} I_{Z_i}$$

- **Remark:** Z_i can be dependent, and I_Z can be much smaller.

Range Estimation

- **Goal: To "encapsulate" each epistemic variable**

- **Overwhelming probability condition:**

For each $I_{Z_i} = \left[\alpha_i, \beta_i\right], \quad \alpha_i < \beta_i,$ find a bounded interval

$$I_{X_i} = [a_i, b_i], \qquad -\infty < a_i < b_i < \infty,$$

such that

$$\Pr(Z_i \in I_i^-) \le \delta_i,$$

where $\delta_i \ge 0,$ and I_i^- is the difference set

$$I_i^- = I_{Z_i} \,\Delta\, I_{X_i} = \left(I_{Z_i} \cup I_{X_i}\right) \backslash \left(I_{Z_i} \cap I_{X_i}\right)$$

- If Z_i is bounded, it is "easier" to do
- If Z_i is unbounded, X_i needs to be "big" enough

Range Estimation (cont'd)

- For each variable:

$$I_{Z_i} = [\alpha_i, \beta_i], \quad -\infty \le \alpha_i < \beta_i \le \infty$$

$$I_{X_i} = [a_i, b_i], \quad -\infty < a_i < b_i < \infty$$

$$\Pr(Z_i \in I_i^-) \le \delta$$

- For all variables:

$$I_Z \subseteq \mathop{\times}_{i=1}^{d} I_{Z_i}$$

$$I_X = \mathop{\times}_{i=1}^{d} I_{X_i} = \mathop{\times}_{i=1}^{d} [a_i, b_i]$$

$$I^+ = I_Z \cup I_X, \qquad I^o = I_Z \cap I_X$$

$$I^- = I_Z \,\Delta\, I_X = I^+ \setminus I^o \ \text{(difference set)}$$

- Overwhelming probability condition:

$$\Pr(Z_i \in I^-) \le \delta, \qquad \delta = 1 - (1 - \delta_i)^d$$

- Reminder: I_X may not overlap I_Z

Encapsulation Problem

- **Original Problem:**

$$\begin{cases} \dfrac{\partial v}{\partial t}(t,x,Z) = \mathcal{L}(v), & D \times (0,T] \times I_Z \\ \mathcal{B}(v) = 0, & \partial D \times [0,T] \times I_Z \\ v = v_0, & D \times \{t=0\} \times I_Z \end{cases}$$

- **Encapsulation Problem:**

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x,X) = \mathcal{L}(u), & D \times (0,T] \times I_X \\ \mathcal{B}(u) = 0, & \partial D \times [0,T] \times I_X \\ u = v_0, & D \times \{t=0\} \times I_X \end{cases}$$

○ Solution in a hypercube: $u(x,t,X) : \bar{D} \times [0,T] \times I_X \to \mathbb{R}$

$$I_X = [a_i, b_i]^d \, (= [-1,1]^d, = [0,1]^d)$$

- **Assumption:** $u(\cdot,\xi) = v(\cdot,\xi), \quad \forall \xi \in I^o$

Solution Strategy of the Encapsulation Problem

- **Encapsulation Problem:**

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x,X) = \mathcal{L}(u), & D \times (0,T] \times I_X \\ \mathcal{B}(u) = 0, & \partial D \times [0,T] \times I_X \\ u = v_0, & D \times \{t=0\} \times I_X \end{cases}$$

- **Solution strategy:** **<u>Controllability on point-wise error</u>**

$$\varepsilon_n = \| u - u_n \|_{L^\infty(I_X)} \to 0, \qquad n \to \infty$$

  - ○ u_n is a good approximation in the entire domain I_X (hypercube)
  - ○ Can "sample" u_n accurately for all realizations
  - ○ No probability distribution is assigned in I_X.
  - ○ Convergence is a mathematical preference, not a practical necessity

- Requirement on error control is strong but achievable
  - ○ Sparse grid collocation (with sufficient regularity)
  - ○ Most stochastic Galerkin methods are possible
  - ○ Without sufficient regularity --- multi-element collocation

Solution “Statistics”

- Solution of the original problem:

$$v(\ ;Z): I_Z \to \mathbb{R} \qquad \mu = \int_{I_Z} v(s)\rho_Z(s)ds$$

- Solution in the hypercube: $I^o = I_Z \cap I_X$

$$u_n(\ ;X): I_X \to \mathbb{R} \qquad \mu_n = \int_{I^o} u_n(s)\rho_Z(s)ds$$

Theorem: Assume $v(Z)$ is bounded and let $C_v = \|v\|_{L^\infty(I_Z)}$. Let u_n be an approximation to the solution of the encapsulation problem $u(X)$, s.t.,

$$\varepsilon_n = \|u - u_n\|_{L^\infty(I_X)}.$$

Then the approximation of the mean solution satisfies

$$|\mu - \mu_n| \le \varepsilon_n + C_v \cdot \delta$$

Mixed Aleatory and Epistemic Uncertainty

- **Original Problem:**

$$\begin{cases} \dfrac{\partial v}{\partial t}(t,x,Y,Z) = \mathcal{L}(v), & D\times(0,T]\times I_Y \times I_Z \\ \mathcal{B}(v)=0, & \partial D\times[0,T]\times I_Y\times I_Z \\ v = v_0, & D\times\{t=0\}\times I_Y\times I_Z \end{cases}$$

Y: random variables with known distribution $F_Y(y) = \Pr(Y \le y), \quad y \in \mathbb{R}^{n_Y}$

Z: n_Z random variables with unknown distribution

- **Two encapsulation choices:** I_{X_i} encapsulates I_{Z_i} with overwhelming probability
  - ○ Separate construction: $I_X = \times_{i=1}^{d} I_{X_i} = \times_{i=1}^{d} [a_i, b_i]$

$$u_n(X,Y) = \hat{u}_m(X) \otimes \tilde{u}_k(Y)$$

  - ○ Simultaneous construction:

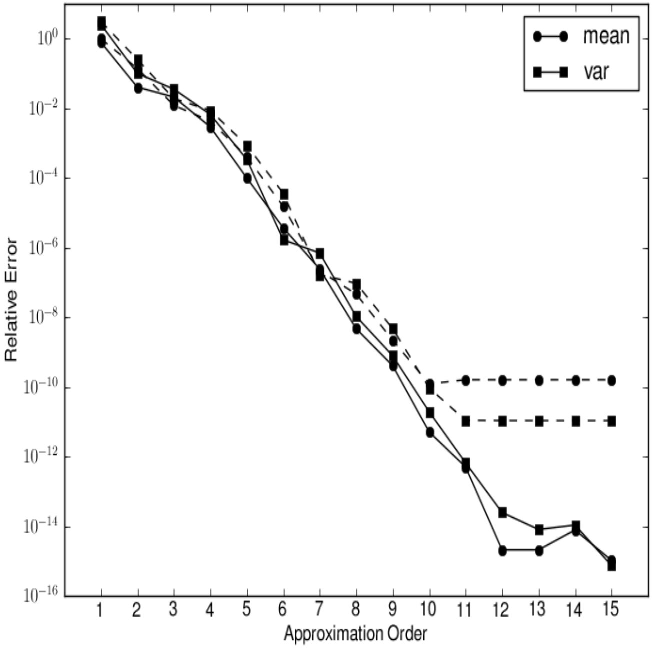
$$I_X = \left(\times_{i=1}^{N_Y} I_{Y_i}\right) \times \left(\times_{i=1}^{N_Z} I_{X_i}\right)$$

Numerical Examples: ODE

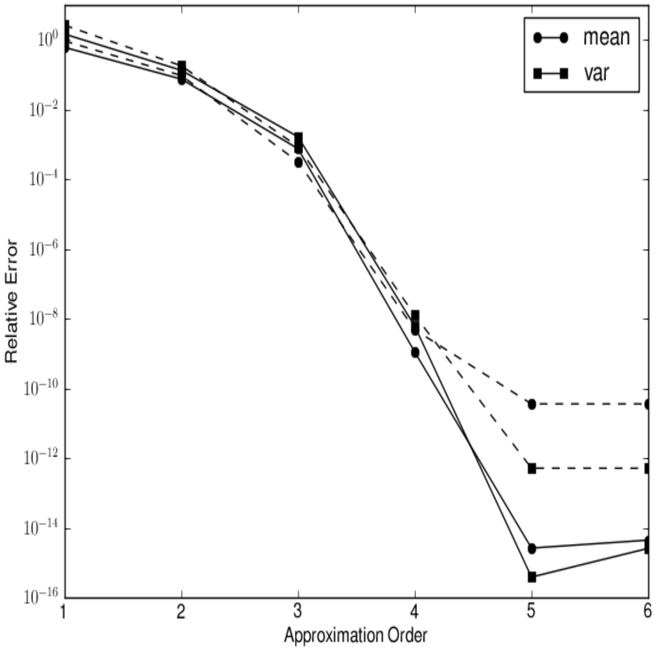
- **Original Problem:** $\frac{dv}{dt}(t) = -Z_1 v, \qquad v(0) = Z_2 \qquad v(t,Z) = Z_2 \exp(-Z_1 t)$
- **Encapsulation Problem:** $\frac{du}{dt}(t) = -X_1 u, \qquad u(0) = X_2, \qquad X = (X_1, X_2) \in [-1,1]^2$

$Z_1, Z_2 \sim$ beta(0,1,1,1), i.i.d. (solid lines) $\qquad I_Z = [0,1]^2$

$Z_1 \sim$ beta(0,1,2,5), $Z_2 \sim$ beta(0,1,1,1) independent (dashed lines)



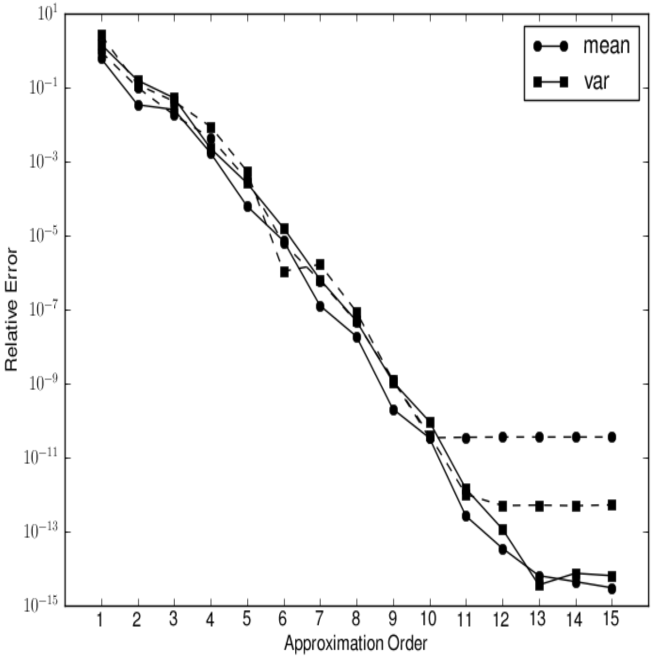
Galerkin with Legendre basis



Sparse grid collocation

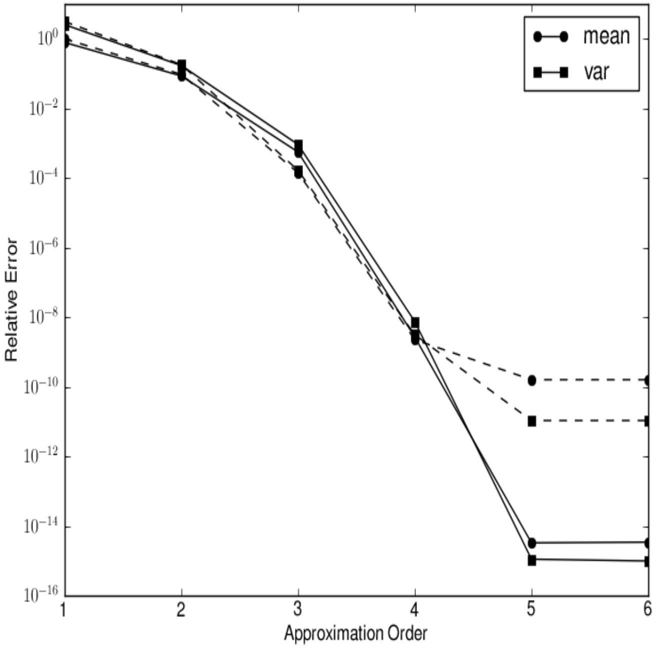
- Z_1 and Z_2 are dependent. In fact, let Z_2=Z_1. Then I_Z=[0,1]

$Z_1 = Z_2 \sim \text{beta}(0,1,1,1)$ (solid lines)



Galerkin with Legendre basis

$Z_1 = Z_2 \sim \text{beta}(0,1,2,5)$ (solid lines)



Sparse grid collocation

I_X=[-1,1]^2

A Slighted More Complicated Example

- **Original Problem:**

$$\frac{d^2 v}{dt^2}(t,Z)+\gamma\frac{dv}{dt}+kv=f\cos(\omega t),\qquad v(0)=v_0,\quad \frac{dv}{dt}(0)=v_1$$

$$Z=(\gamma,k,f,\omega,v_0,v_1)\in\mathbb{R}^6$$

- **Encapsulation Problem:**

$$\frac{d^2 u}{dt^2}(t,X)+X_1\frac{du}{dt}+X_2 u=X_3\cos(X_4 t),\qquad u(0)=X_5,\quad \frac{du}{dt}(0)=X_6$$

$$X=(X_1,\ldots,X_6)\in[-1,1]^6$$

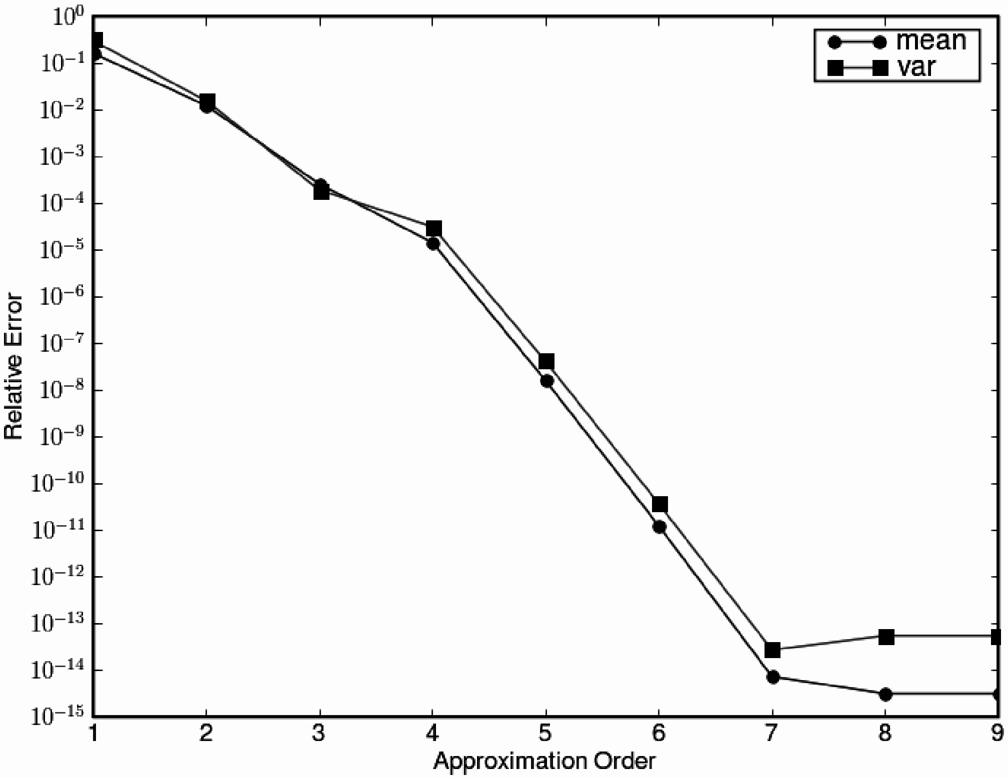
○ Solved by 6-dimensional sparse grid collocation for $t=20$

Dependent Inputs

$$\frac{d^2u}{dt^2}(t,Z)+Z_1\frac{du}{dt}+Z_2u=Z_3\cos(Z_4t), \qquad u(0)=Z_5, \quad \frac{du}{dt}(0)=Z_6$$

$Z_1 \sim \text{beta}(0.08,0.12,3,2)$, $Z_3 \sim \text{beta}(0.08,0.1,1,1)$, $Z_5 \sim \text{uniform}(0.45,0.55)$, independent

$$Z_2=Z_1^2/4+0.01,\ Z_4=10Z_3,\ Z_6=(Z_5-0.5)$$

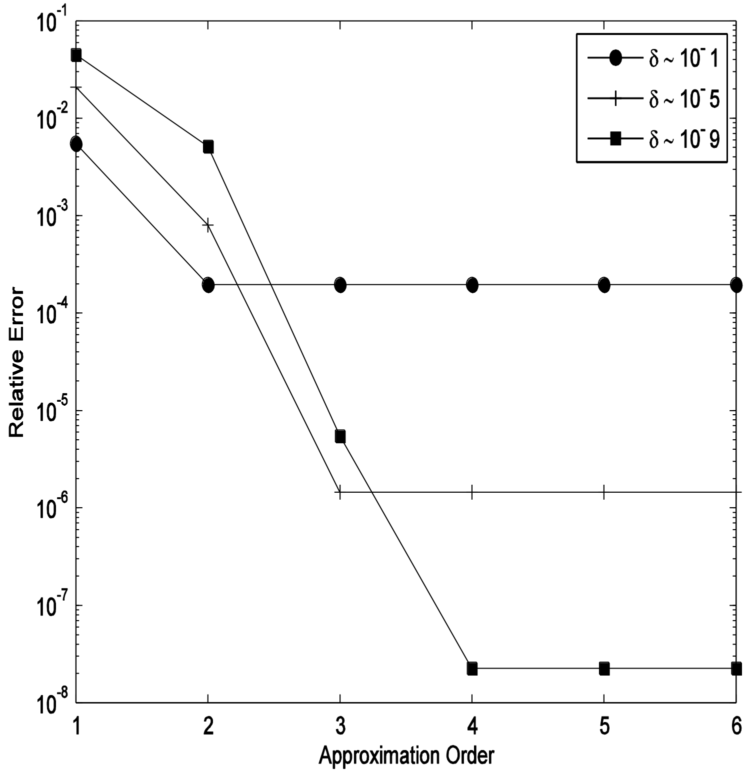


Unbounded Inputs

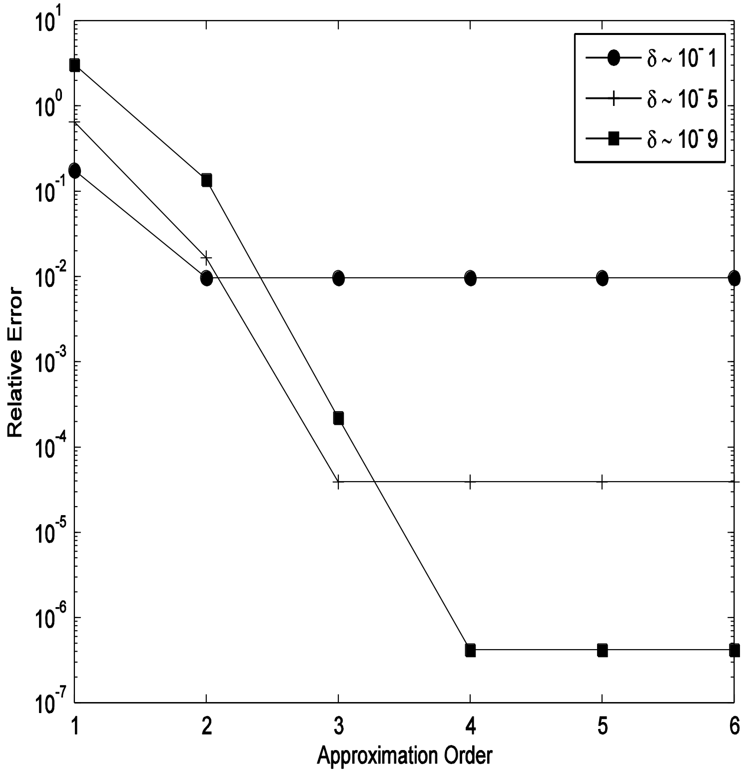
$$\frac{d^2u}{dt^2}(t,Z)+Z_1\frac{du}{dt}+Z_2u=Z_3\cos(Z_4t), \qquad u(0)=Z_5, \quad \frac{du}{dt}(0)=Z_6$$

Gaussian: $Z\sim\mathcal{N}(\mathbf{0},\mathbf{C})$, $\mathbf{C}\in\mathbb{R}^{6\times 6}$ is the covariance matrix

$$I_Z=\mathbb{R}^6, \quad I_X=[-a,a]^6, \quad \text{then } \delta>0$$



Mean

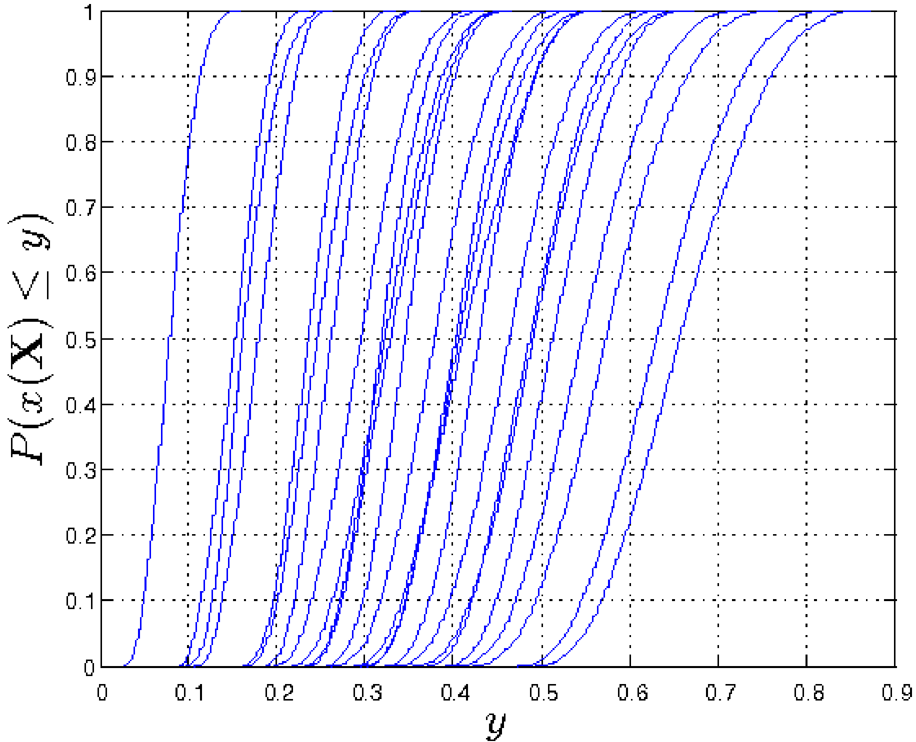


Variance

Mixed Aleatory and Epistemic Case

$$\frac{d^2u}{dt^2}(t,Z) + Z_1\frac{du}{dt} + Z_2 u = Z_3\cos(Z_4 t), \qquad u(0) = Z_5, \quad \frac{du}{dt}(0) = Z_6$$

- Aleatory: $Z_1 \sim \text{beta}(0,1,0,0),\ Z_2 = Z_1^2/4 + 0.01, Z_3 \sim \text{beta}(0,1,1,1),\ Z_5 \sim \text{beta}(0,1,2,1)$
- Epistemic: $Z_4 \in [0.8, 1.2], \quad Z_6 \in [-0.05, 0.05]$



CDF of the aleatory RVs at 25 prescribed values of the epistemic variables

- Solution obtained by sparse grid collocation via simultaneous construction (5-d)

Stochastic Diffusion Equation

$$-\frac{d}{dx}\left[a(x,Z)\frac{dv}{dx}(x,Z)\right]=f,\qquad (x,Z)\in(0,1)\times I_Z$$

$$a(x,Z)=1+\sigma\sum_{k=1}^{d}\frac{1}{k^2\pi^2}\cos(2\pi kx)Z_k$$

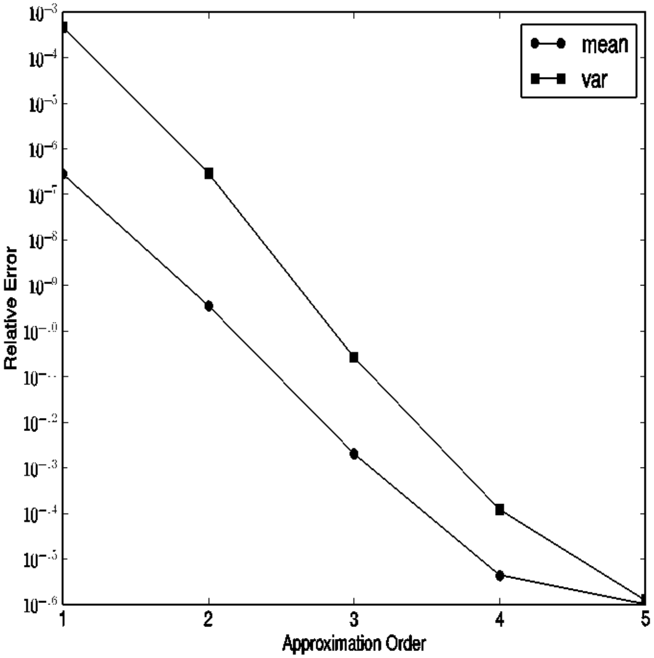
- All *Z*’s are epistemic $\sigma = 4$

$$Z_1 \sim \mathrm{beta}(0,1,3,2),\quad Z_3 \sim \mathrm{beta}(-1,0,1,1),\quad Z_5 \sim \mathrm{beta}(-0.5,0.5,0,0)$$

$$Z_2 = Z_1 Z_5,\quad Z_4 = (Z_1^2+1)Z_3,\quad Z_6 = -Z_5$$

- Solution obtained by Legendre-Galerkin

$$X \in [-1,1]^6$$



Data Driven Uncertainty Analysis

- Experimental data:
  - Expensive, inflexible → ***USE THEM!***
  - Scarce, incomplete

- Measurements:
  - System parameters → to characterize inputs
  - Solutions → to further refine predictions
  - Other indirect quantities → can still be helpful

Data: Inverse Parameter Estimation

- **Stochastic PDE:**

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x,Z) = \mathcal{L}(u), & (0,T]\times D\times \mathbb{R}^{n_z} \\ \mathcal{B}(u) = 0, & [0,T]\times \partial D\times \mathbb{R}^{n_z} \\ u = u_0(x,Z), & \{t=0\}\times D\times \mathbb{R}^{n_z} \end{cases}$$

- **Solution:**

$$u(t,x,Z):[0,T]\times \overline{D}\times \mathbb{R}^{n_z} \mapsto \mathbb{R}^{n_u}$$

- **Prior distribution:**

$$\pi_Z(z) = \prod_{i=1}^{n_z} \pi_i(z_i)$$

- Estimation of the prior distribution
  - Requires direct measurements of the parameters
  - No/not enough direct measurements? (Use experience/intuition …)
  - How to take advantage of measurements of other variables?

Bayesian Inference for Parameter Estimation

- **Available information:**
    - Prior distribution of the parameters *Z* (based on assumptions)
    - Simulation of measurable variables: *G(Z) – forward problem*
    - Measurement of computable variables --- data *d*

$$d = G(Z) + e, \quad e \in \mathbb{R}^{n_d} \text{ is i.i.d.}$$

- **Goal:** To estimation the distribution of *Z* --- posterior distribution

- **Posterior distribution:** $\pi(Z\,|\,d) \propto \pi(d\,|\,Z)\pi(Z)$

- **Likelihood function:** $\pi(d\,|\,Z) = \prod_{i=1}^{n_d} \pi_{e_i}(d_i - G_i(Z))$

- **Notes:**
    - Difficult to manipulate
    - Classical sampling approaches can be time consuming (MCMC, etc)

gPC Based Algorithm

$$\pi(Z \mid d) \propto \pi(d \mid Z)\pi(Z) \qquad \pi(d \mid Z) = \prod_{i=1}^{n_d} \pi_{e_i}(d_i - G_i(Z))$$

- **Fast gPC algorithm:** *(Marzouk, Najm, Rahn, JCP, 07)*
  - Replace the forward problem G with a gPC solution G_N
  - Sampling of the gPC solution for an arbitrarily large number of times

$$\pi_N(Z \mid d) \propto \pi_N(d \mid Z)\pi(Z) \qquad \pi_N(d \mid Z) = \prod_{i=1}^{n_d} \pi_{e_i}(d_i - G_{N,i}(Z))$$

- **Properties:**
  - Allows direct sampling in term of Z with arbitrarily large samples
  - No additional simulations – forward problem solver only ("Free lunch")

Convergence of gPC Bayesian Inference

- **Kullback-Leibler divergence:** $D\left(\pi_1 \| \pi_2\right) \triangleq \int \pi_1(z) \log \frac{\pi_1(z)}{\pi_2(z)} dz$

- **Observation error:** $e \sim N(0, \sigma^2 \mathbb{I})$, i.i.d. Normal

Theorem. If the gPC expansion G_N converges to G in $L^2_{\pi_Z}$, then the posterior density π^d_N converges to π^d in the sense

$$\left| \mathrm{D}\left(\pi^d_N \| \pi^d\right) \right| \to 0, \quad N \to \infty.$$

Moreover, if

$$\left\| G_{\mathrm{i}}(Z) - G_{N,i}(Z) \right\|_{L^2_{\pi_Z}} \le C N^{-\alpha}, \quad 1 \le i \le n_d, \alpha > 0, \ C \text{ independent of } N,$$

then for sufficiently large N,

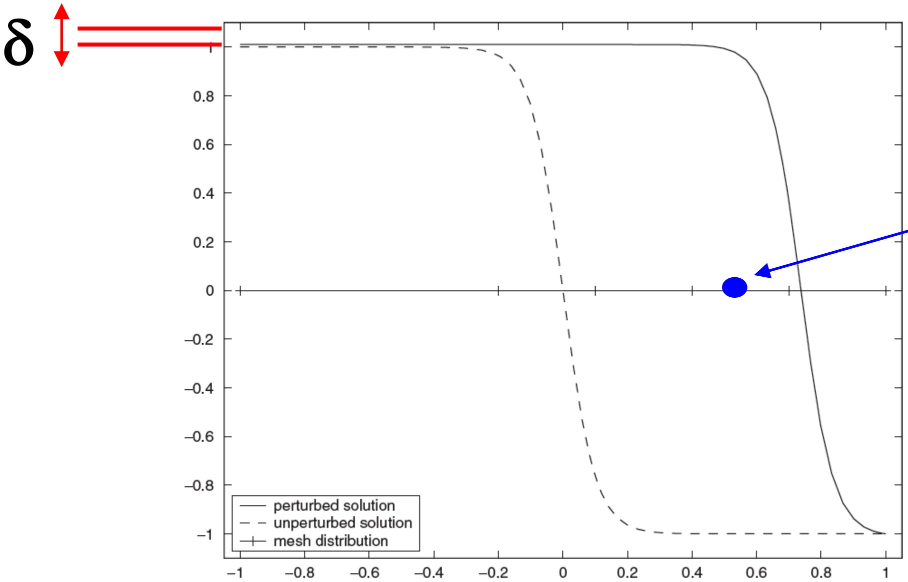
$$\left| D\left(\pi^d_N \| \pi^d\right) \right| \lesssim N^{-\alpha}.$$

- **Notes:**
  - Fast (exponential) convergence rate is retained

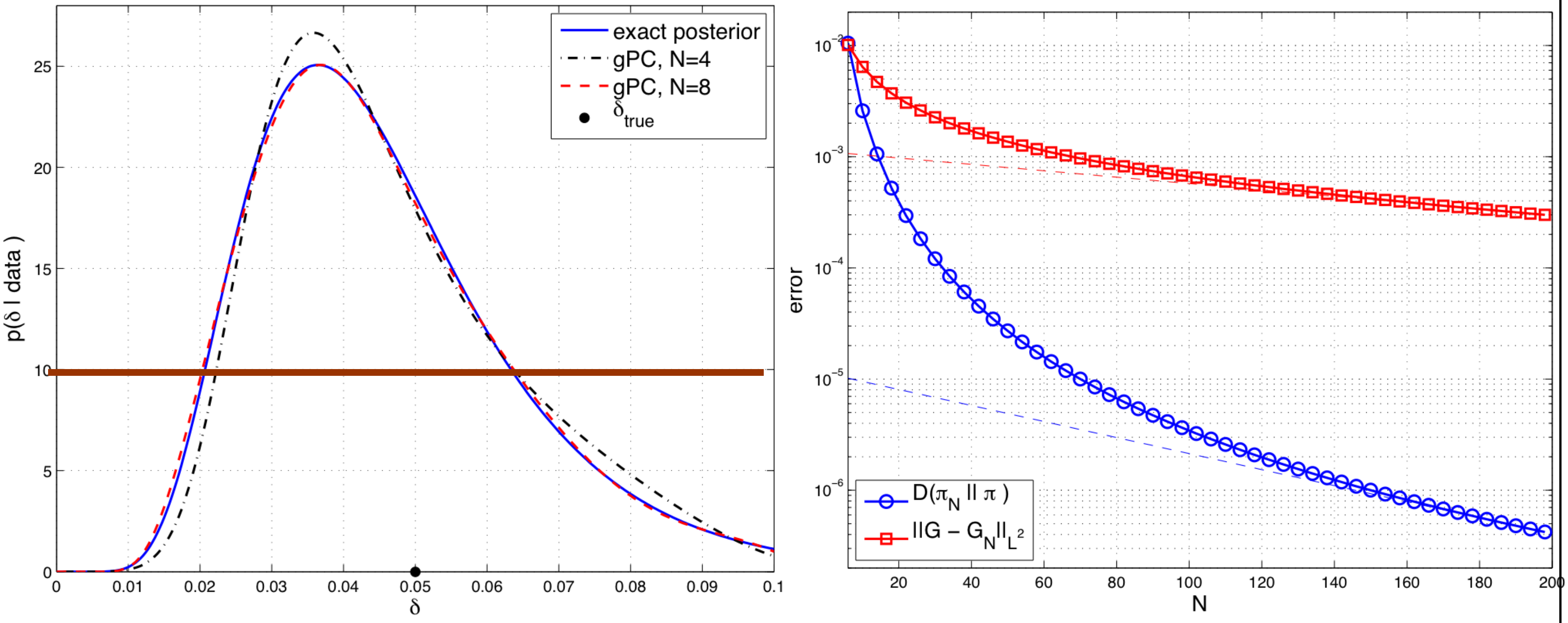
(Marzouk & Xiu, Comm. Comput. Phys.,vol. 6, 08)

Parameter Estimation: Supersensitivity Example

- Burgers' equation : $\dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = \nu \dfrac{\partial^2 u}{\partial x^2}, \quad x \in -1,1$
- Boundary conditions : $u(-1) = 1 + \delta(Z); \quad u(1) = -1; \quad 0 < \delta << 1$



noisy observation of transition layer location

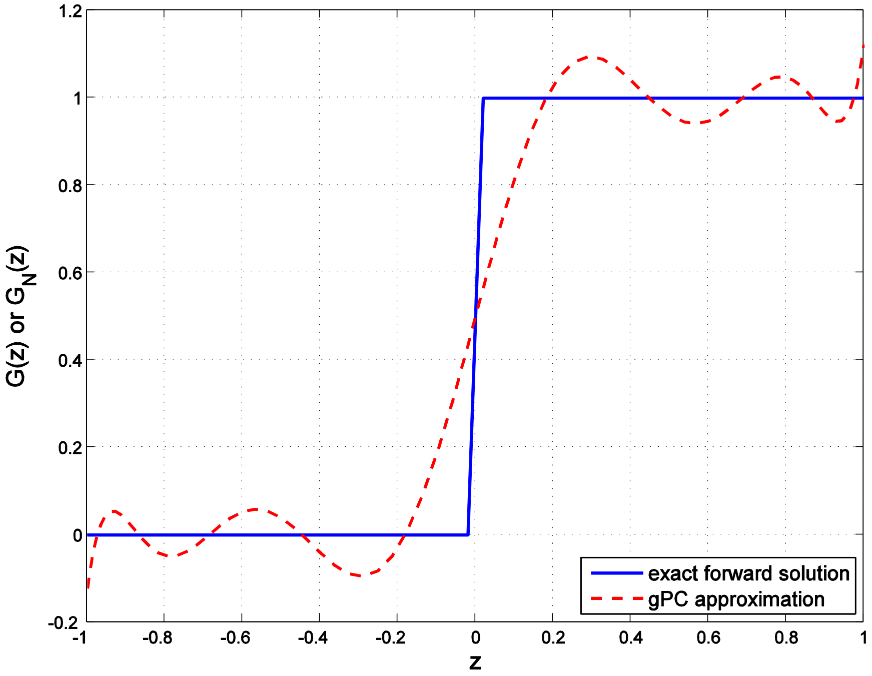


Prior distribution is uniform *Z~(0,0.1)*

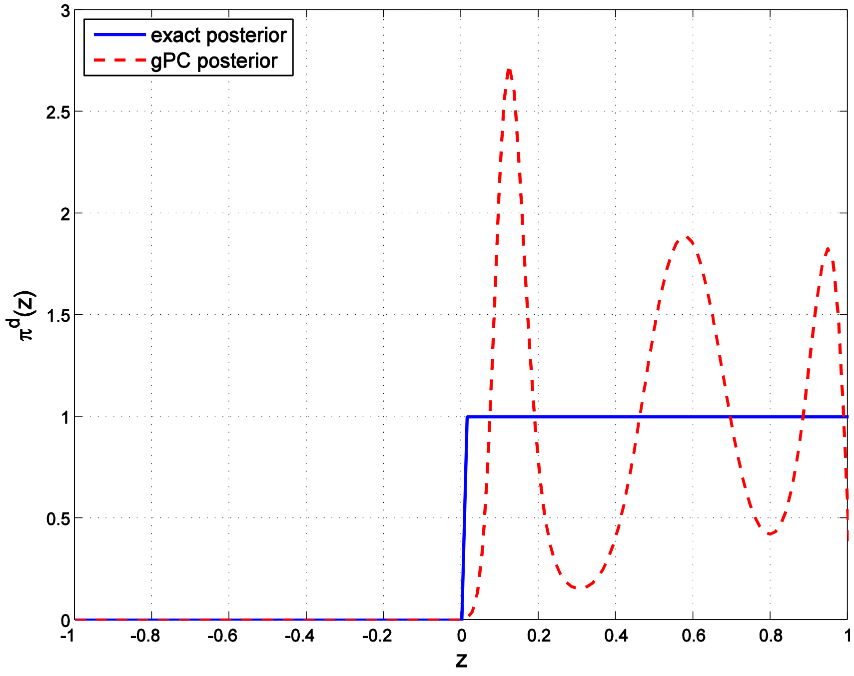
Measurement noise: $e \sim N(0, 0.05^2)$

Parameter Estimation: Step Function

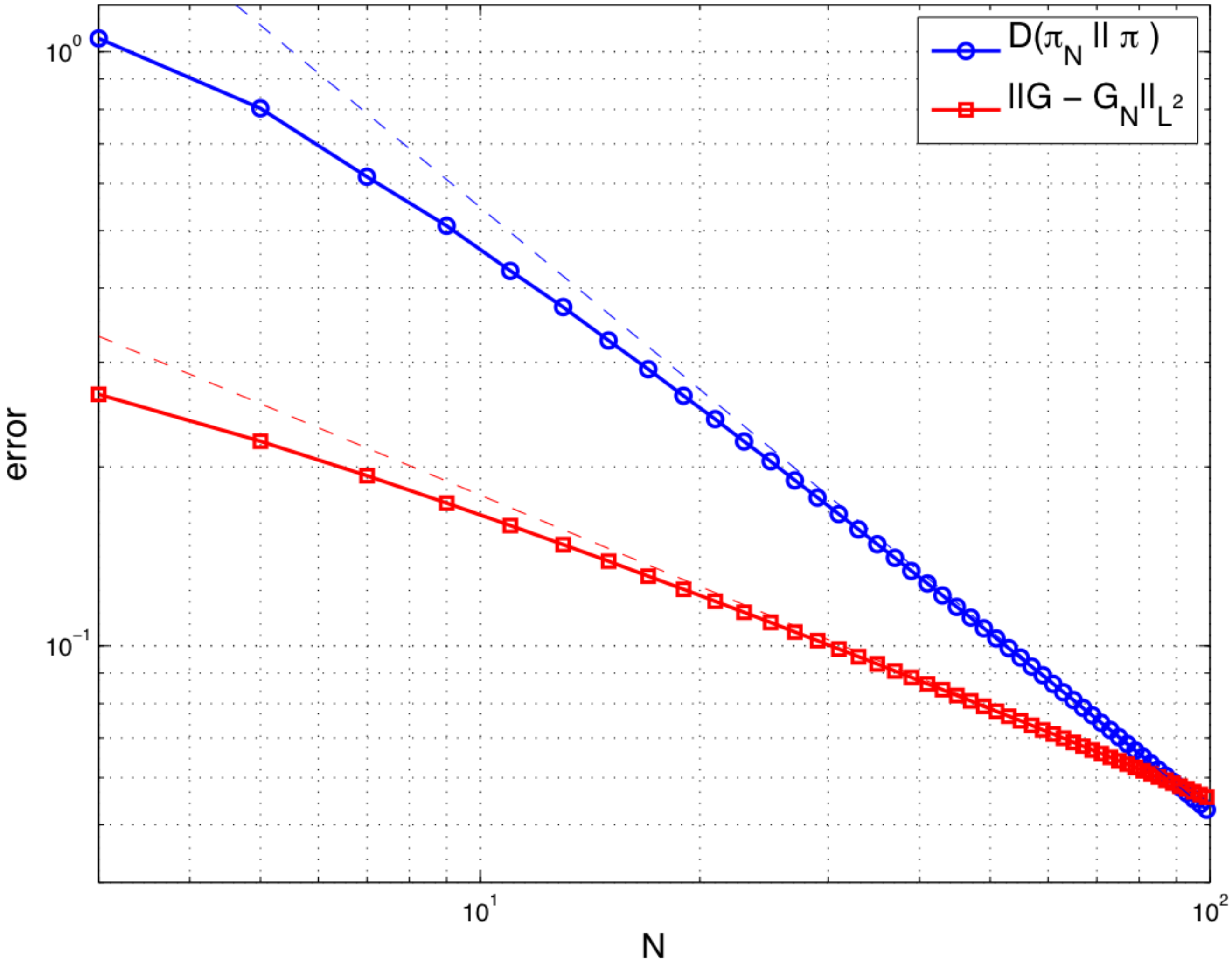
- Assume the forward model is a step function
- Posterior distribution is discontinuous
- Gibb’s oscillations exist
- Slow convergence with global gPC basis functions



Forward model and its approximation



Posterior distribution and its approximation



$D(\pi_N \| \pi)$
$\|G - G_N\|_{L^2}$
error
N
10^0
10^{-1}
10^1
10^2

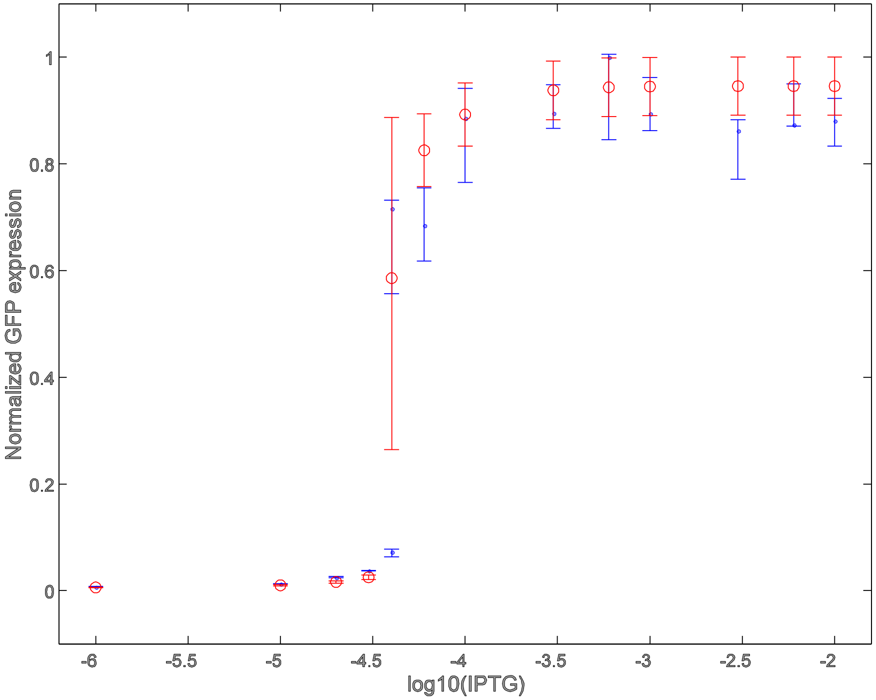
Stochastic collocation

- **Example:** estimate kinetic parameters in a *genetic toggle switch*
  - Differential-algebraic equations model from [Gardner *et al*, Nature, 2000]
  - **Real experimental data**: steady-state expression levels of one gene (v)

$$\frac{du}{dt} = \frac{\alpha_1}{1+v^{\beta}} - u$$

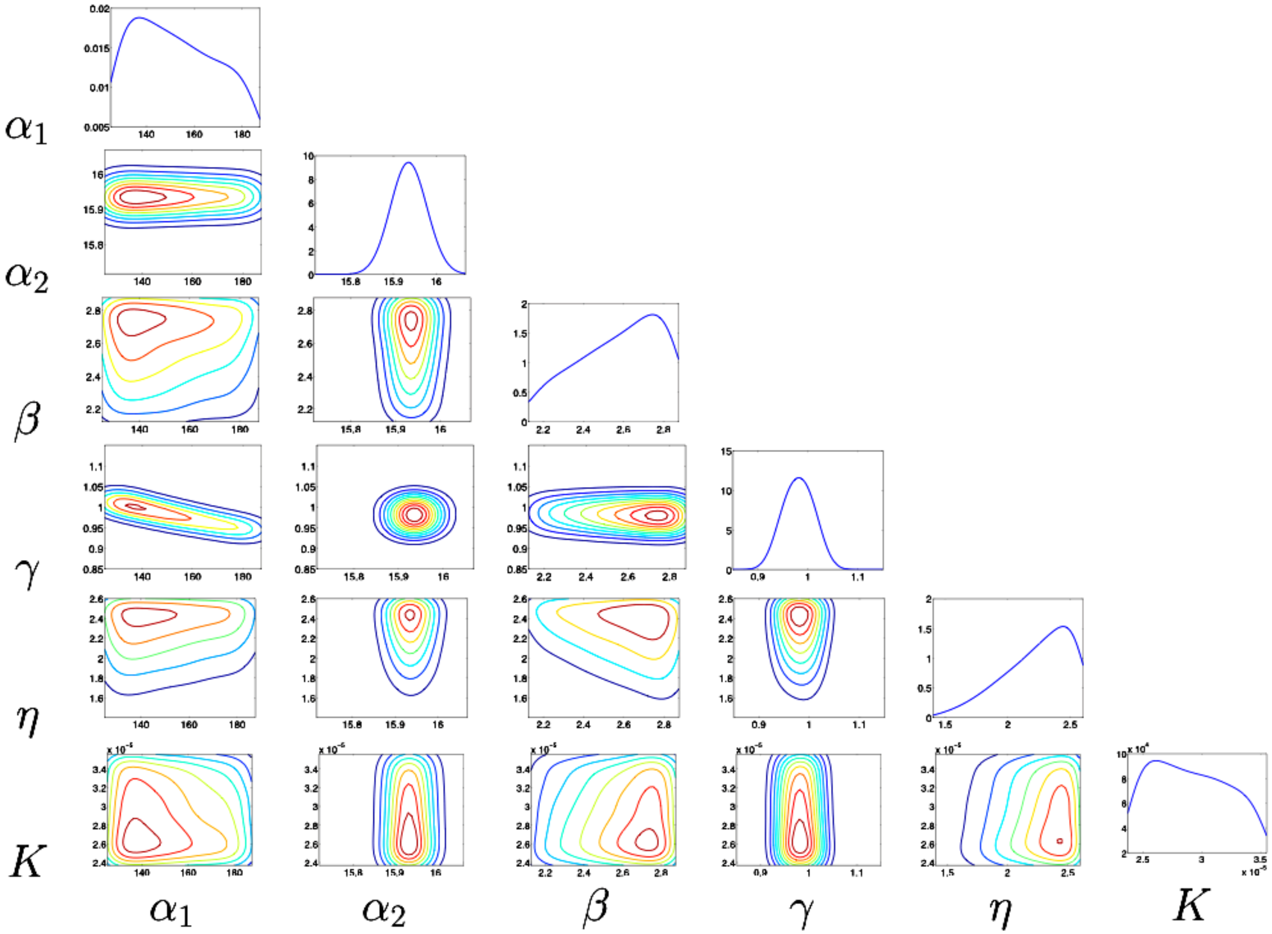
$$\frac{dv}{dt} = \frac{\alpha_2}{1+w^{\gamma}} - v$$

$$w = \frac{u}{(1+[IPTG]/K)^{\eta}}$$



- 6 uncertain parameters
- Assumed to be independent and uniform in the forward problem
- Good agreement with measurement --- but let’s now use the data *again*

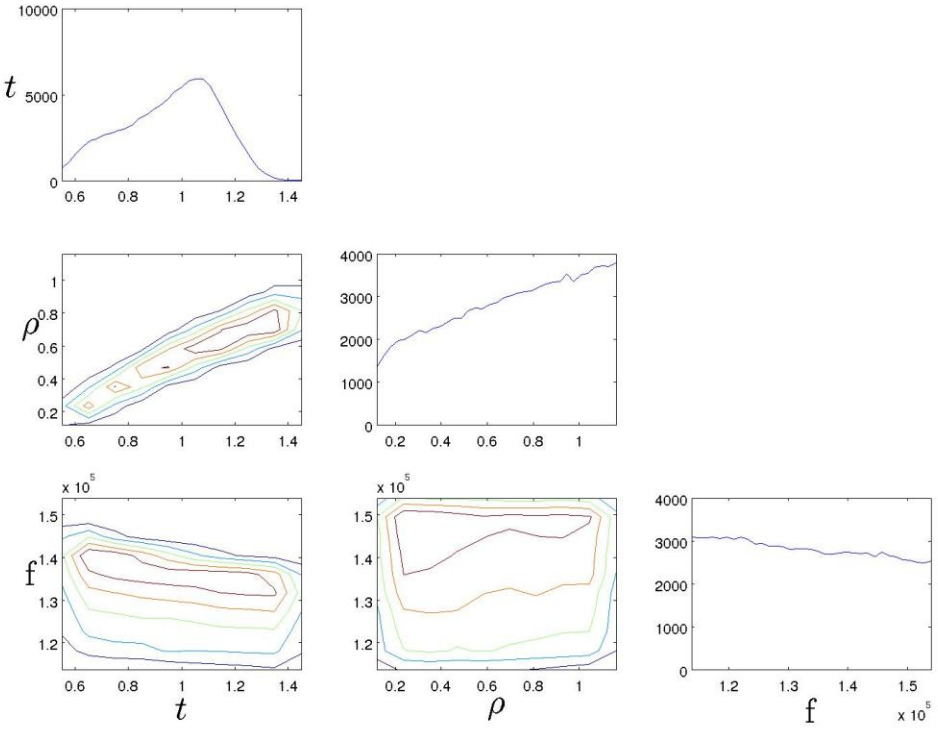
Using the Data Twice – Parameter estimation

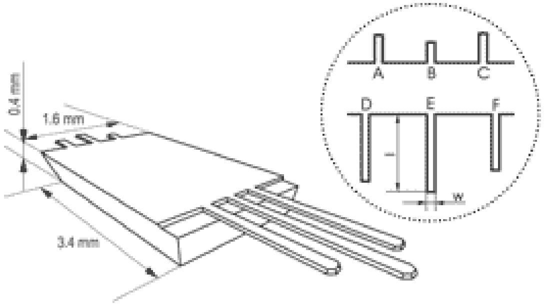


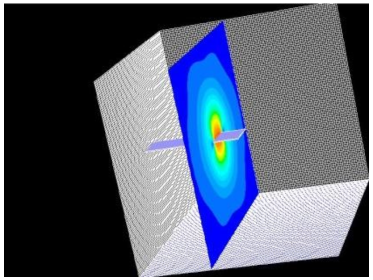
1-parameter and 2-parameter marginal posterior densities
(*dim* = 6, *N* = 4, 6-level **sparse grid** forward problem solver)

Bayesian Inverse Estimation for Free-Cantilever Damping

Gas damping of freely-vibrating cantilever

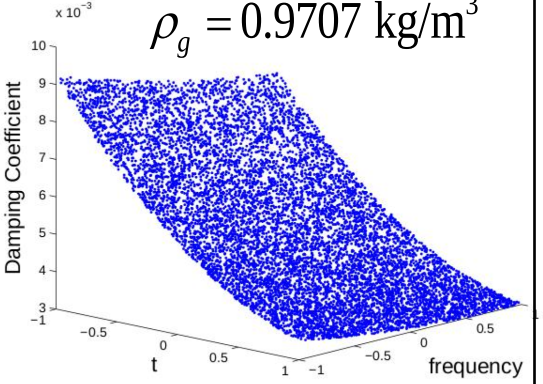






MEMOSA for damping simulation

$\rho_g = 0.9707\ \mathrm{kg/m^3}$



3-parameter marginal posterior densities
Level 2 sparse grid in thickness, gas density and frequency

Summary

- Uncertainty Analysis: To provide improved prediction
  - Input characterization
  - Uncertainty propagation
  - Post processing

- Generalized polynomial chaos (gPC)
  - Multivariate approximation theory

- The *“Freebies”* of gPC
  - Parameter estimation
  - Data assimilation: Kalman filter
  - Distributional sensitivity analsysis
  - Epistemic uncertainty

- **Data, *any* data, could help**