SAND2015-1887C

Accounting for Model Error in the Calibration of Physical Models

H. Najm[1], K. Sargsyan[1], J. Bender[1], and R. Ghanem[2]

[1]Sandia National Laboratories
Livermore, CA
[2]University of Southern California
Los Angeles, CA

SIAM CS&E
Salt Lake City, UT
Mar 14-18, 2015

Acknowledgement

B.J. Debusschere, R.D. Berry, C. Safta, K. Chowdhary
— Sandia National Laboratories, CA

O.M. Knio — Duke Univ., Durham, NC, USA
O.P. Le Maître — CNRS, Paris, France
Y.M. Marzouk — Mass. Inst. of Tech., Cambridge, MA, USA

This work was supported by:

- DOE Office of Basic Energy Sciences, Div. of Chem. Sci., Geosci., & Biosci.
- DOE Office of Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC)

Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

Outline

Motivation

- All models are wrong in principle
- Models of physical systems rely on
  - Presumed theoretical framework
  - Mathematical formulation
- Practical models of complex physical systems rely on
  - Simplifying assumptions
  - Numerical discretization of governing equations
  - Computational software & hardware
- model error is frequently non-negligible
- Estimating model error is useful for
  - model comparison & validation
  - model improvement & scientific discovery
  - reliable computational predictions

Statistical modeling of model error

Error framework:

Measurements: $y_{\text{data}} = y_{\text{truth}} + \epsilon_d$

Model predictions: $y_{\text{truth}} = y_{\text{model}} + \epsilon_m$

Thus: $y_{\text{data}} = y_{\text{model}} + \epsilon_m + \epsilon_d$

Error modeling – example

Model: $y_{\text{model}} = f(x, \lambda)$

Data Error: $\epsilon_d \sim \text{N}(0, \sigma^2)$

Model Error: $\epsilon_m \sim \text{GP}(\mu(x), C(x, x'))$

Model calibration:

Estimate model parameters λ along with those of ϵ_m, ϵ_d

Challenges – Physical Models

- arbitrary choice of statistical model (*e.g.* GP) spatial structure does not take the physical model into acct
  - potential violation of implicit constraints in physical models
  - *e.g.* incompressible flow: $\nabla \cdot v = 0$
- difficulty in disambiguation of model & data error
- calibration of model error on measured observable does not impact quality of other model predictions
- Physical scientists are unlikely to augment their model with a statistical model error term on select outputs

Key idea - Targeted model error embedding

- Embed model error in specific submodel phenomenology (Berliner)
  - a modified transport or constitutive law
  - a modified formulation for a material property
- Pros:
  - Allows placement of model error term in locations where key modeling assumptions and approximations are made
    - as a correction or high-order term
    - as a possible alternate phenomenology
  - explore if it can explain discrepancy on observable
  - naturally preserves model structure and associated constraints
- Cons:
  - complex likelihood $p(y|\lambda)$ for general nonlinear $f(x, \lambda, \epsilon_m)$

Consider a simple no-data-noise setting

- Calibration of a (simple) model against a complex model
- Let the complex model be presumed to represent the truth
- In this context, the data has no noise
- Discrepancy between model and data is all due to model error

$$y_{\text{data}} = y_{\text{truth}} = y_{\text{complex\_model}} = y_{\text{model}} + \epsilon_m$$

- $\epsilon_m = y_{\text{data}} - y_{\text{model}}$ is a deterministic quantity
- The only information as to the quality of the calibrated uncertain model, *e.g.* via a posterior predictive check, is in a unique ϵ_m for any x

model-to-model calibration

Model: $y = f(x, \lambda, \phi(\epsilon_m))$

– Random variable ϕ in augmented model components carries model error

Data: $D = \{(x_i, y_{\text{data},i}), i = 1, \ldots, N\}$

- Goal:
  - Establish λ, $p(\phi)$ such that the likelihood of the data is high, based on the posterior predictive $p(y|D)$
- This puts us in a density estimation framework for ϕ:
  - The utility of additional data is to improve the specification of λ, and $p(\phi)$

Present Context

Embed ϵ_m in λ

- Model: $y = f(x, \lambda)$ with $\lambda : \Omega \to \mathbb{R}^M$
- Density estimation problem for $p(\lambda)$
- Let the random variable λ be parameterized by α
  - For example, define λ as a polynomial chaos expansion

$$\lambda = \sum_{k=0}^{P} \alpha_k \Psi_k(\xi)$$

- Parameter estimation problem for $\alpha = (\alpha_0, \cdots, \alpha_P)$
- Bayesian setting
  - Prior $\pi(\alpha)$
  - Likelihood $L(\alpha) = p(D|\alpha)$

Likelihood construction – variants

- Full Likelihood

$$L(\alpha) = p(D|\alpha) = p(y_{\text{data},1}, \ldots, y_{\text{data},N}|\alpha)$$

- Marginalized Likelihood

$$L(\alpha) = p(D|\alpha) = \prod_{i=1}^{N} p(y_{\text{data},i}|\alpha)$$

- Approximate Bayesian Computation (ABC):
  Seek to satisfy the constraints:
  - $p(y|D)$ is “centered" on the data
  - The width of the distribution $p(y|D)$ is “consistent" with the spread of the data around the nominal model prediction

Approximate Bayesian Computation (ABC)

Employ a kernel density as a pseudo-likelihood to enforce select constraints:

- Uncertain prediction $p(y|D)$ is centered on the data
  - With $\mu_i(\alpha) = \mathrm{E}_\xi[f(x_i, \lambda(\xi;\alpha))]$:

    $$\text{minimize } \| \; \mu_i(\alpha) - y_{\text{data},i} \; \|_2^2$$
- The width of the distribution $p(y|D)$ is consistent with the spread of the data around the nominal model prediction
  - With $\sigma_i^2(\alpha) = \mathrm{V}_\xi[f(x_i, \lambda(\xi,\alpha))]$:

    $$\text{minimize } \| \; \sigma_i(\alpha) - \gamma|\mu_i(\alpha) - y_{\text{data},i}| \; \|_2^2$$
  - γ is a factor that specifies the desired match between σ_i and the discrepancy $|\mu_i(\alpha) - y_{\text{data},i}|$, on average

ABC Likelihood

With $\rho(\mathcal{S})$ being a metric of the statistic $\mathcal{S}$, use the kernel function as an ABC likelihood:

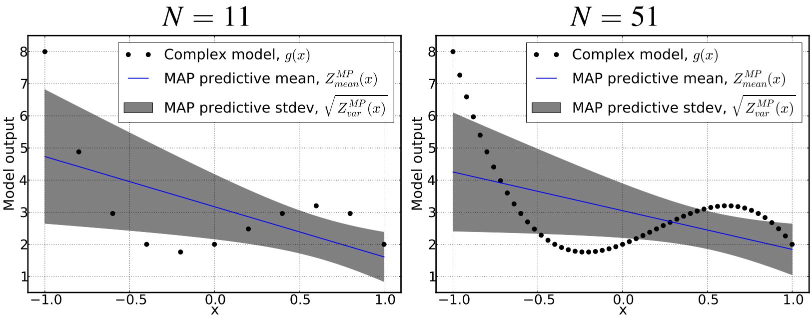
$$L_{\text{ABC}}(\alpha) = \frac{1}{\epsilon} K\left(\frac{\rho(\mathcal{S})}{\epsilon}\right)$$

where ϵ controls the severity of the consistency control

Propose the Gaussian kernel density:

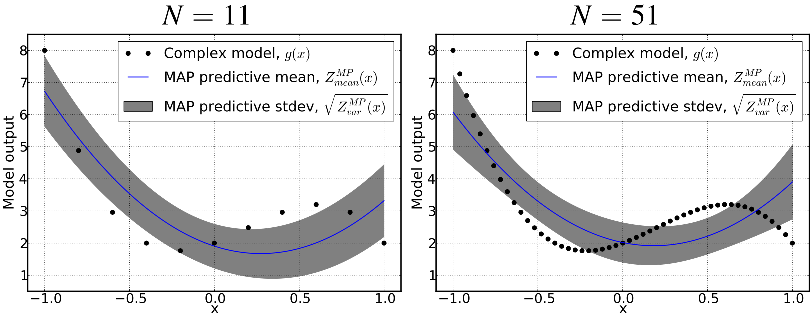
$$L_{\epsilon}(\alpha) = \frac{1}{\epsilon\sqrt{2\pi}} \prod_{i=1}^{N} \exp\left(-\frac{(\mu_i(\alpha) - y_{\text{d},i})^2 + (\sigma_i(\alpha) - \gamma|\mu_i(\alpha) - y_{\text{d},i}|)^2}{2\epsilon^2}\right)$$

Test problem – Cubic data fit by a line



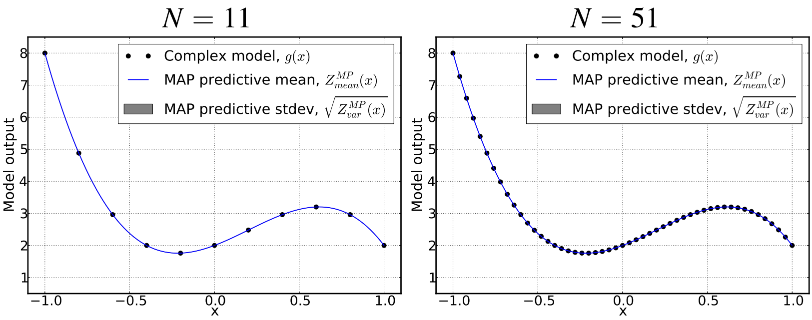
- MAP predictive (MP) mean centered on data
- MP standard deviation captures range of discrepancy
- Increasing number of data points has a small effect on both MP mean and stdev

Test problem – Cubic data fit by a quadratic



- Quadratic has better fit to the data
- Smaller MP stdev consistent with smaller discrepancy

Test problem – Cubic data fit by a cubic



- Cubic has perfect fit to the data
- Negligible MP stdev consistent with negligible discrepancy

Consider a noisy-data setting

- Calibration of a model $y_m = f(x, \lambda)$ against noisy data
- Synthetic noisy data is generated from a “truth" model + Gaussian noise
- Discrepancy between fit model prediction and data is due to both model error & data noise

$$y = y_{\text{data}} = y_{\text{truth}} + \epsilon = f(x, \lambda) + \epsilon$$

- Modeling strategy:
  - Model λ as a random vector, represented with PC
  - Represent the noise similarly using PC
  - Estimate all PC coefficients using ABC

Model Error formulation – noisy data

$$y = f(x, \lambda) + \epsilon$$

Let $\epsilon \sim N(0, \sigma^2)$. With N *i.i.d.* data points we have

$$y_i = f(x_i, \lambda) + \epsilon_i, \quad i = 1, \ldots, N$$

For Hermite-Gaussian PC:

$$\begin{aligned}
\lambda &= \sum_{k=0}^{P} \alpha_k \Psi_k(\xi_1, \cdots, \xi_d), \qquad \alpha \equiv (\alpha_0, \cdots, \alpha_P) \\
f(x, \lambda) &= \sum_{k=0}^{P} f_k(x, \alpha) \Psi_k(\xi_1, \cdots, \xi_d) \\
y_i &= h(x_i, \alpha, \sigma, \xi) = \sum_{k=0}^{P} f_k(x_i, \alpha) \Psi_k(\xi_1, \cdots, \xi_d) + \sigma \xi_{d+i}
\end{aligned}$$

Augmented PC germ $\xi = (\xi_1, \cdots, \xi_d, \xi_{d+1}, \cdots, \xi_{d+N})$

Model Error Estimation – noisy data

Inverse problem:

- Given:
  - data:
    $$D = \{(x_i, y_i)\}_{i=1}^N$$
  - data model:
    for $i = 1, \ldots, N$:
    $$y_i = h(x_i, \alpha, \sigma, \xi) = \sum_k f_k(x_i, \alpha)\Psi_k(\xi_1, \cdots, \xi_d) + \sigma\xi_{d+i}$$
- estimate (α, σ).

Bayesian context:

- posterior: $p(\alpha, \sigma|D)$
- options: Full Bayesian likelihood; Marginalized; ABC
- ABC : Estimate (α, σ) s.t. mean and stdv of $h(x, \alpha, \sigma, \xi)$ are consistent with D in an ℓ_2-sense

Calibrated Uncertain Model Predictions

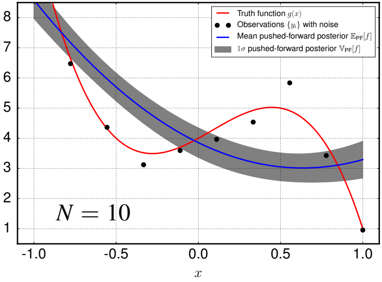
- Calibrated model : $y = f(x;\lambda(\xi;\alpha))$
- Pushed forward posterior : $p(f(x;\lambda(\xi;\alpha))|D)$
- PFP Mean :

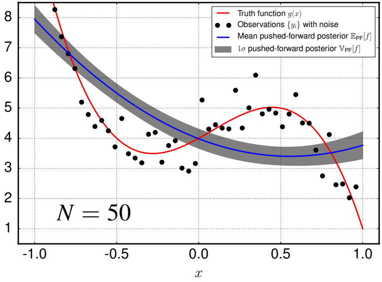
$$\mathbb{E}_{\mathrm{PFP}}[f(x;\lambda(\xi;\alpha))] = \mathbb{E}_\alpha[\mathbb{E}_\xi[f]]$$

- PFP Variance:

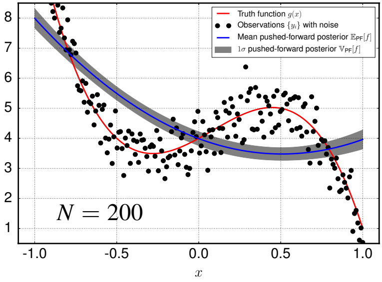
$$\mathbb{V}_{\mathrm{PFP}}[f(x;\lambda(\xi;\alpha))] = \underbrace{\mathbb{E}_\alpha[\mathbb{V}_\xi[f]]}_{\sim \text{model error}} + \underbrace{\mathbb{V}_\alpha[\mathbb{E}_\xi[f]]}_{\sim \text{data noise}}$$

Cubic-quadratic fitting – Classical Bayesian likelihood

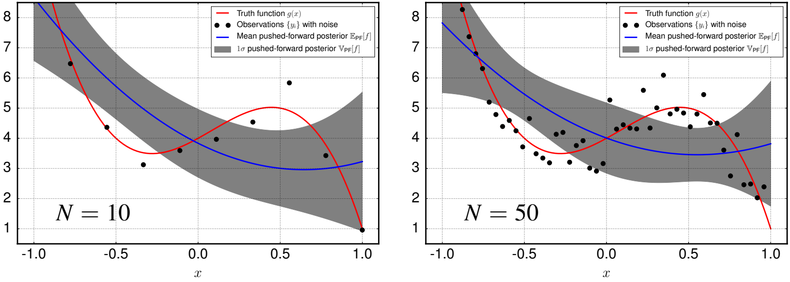




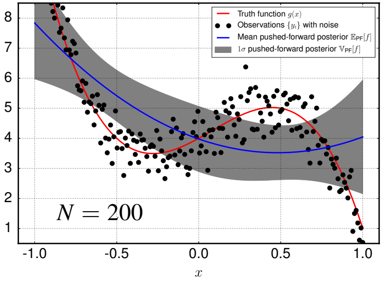
- With additional data, predictive uncertainty around the wrong model is indefinitely reducible
- Predictive uncertainty not indicative of discrepancy from truth



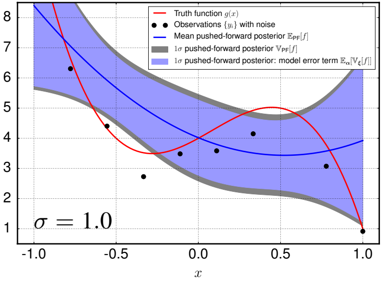
Cubic-quadratic fitting – Model Error ABC Likelihood

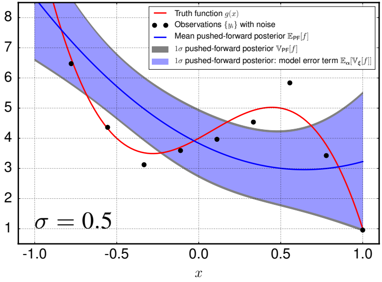


- With additional data, predictive uncertainty around the wrong model is not reducible
- Predictive uncertainty is indicative of discrepancy from truth

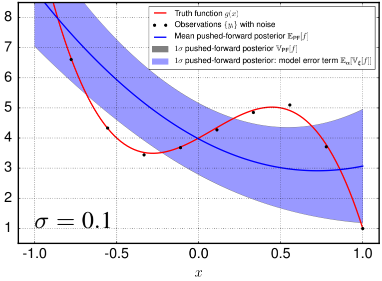


Cubic-quadratic – Model Error – ABC

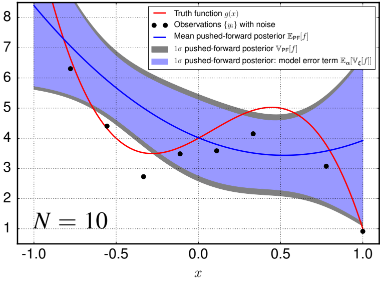


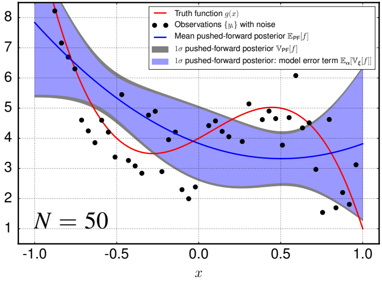


- Predictive uncertainty composed of both model-error and data-noise components
- The data-noise component is reducible with lower-noise in the data

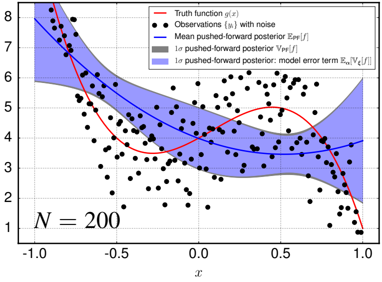


Cubic-quadratic – Model Error – ABC





- With additional data, predictive uncertainty due to data noise is reducible
- Predictive uncertainty due to model error is not reducible



Closure

- Presented a strategy for dealing with model error
    - targeted at physical models
- Density estimation framework – $y = f(x; \lambda(\xi; \alpha))$
- ABC : data model is constrained such that
    - nominal prediction is centered on the data
    - predictive uncertainty is consistent with discrepancy fr. data
- Uncertain predictions with the calibrated model include uncertainty due to both model-error and data-noise
- Results suggest disambiguation of the two components
- Uncertainty due to data-noise:
    - Manifested in $\mathbb{V}_\alpha[\mathbb{E}_\xi[f]]$ – Reducible with more/cleaner data
- Uncertainty due to model-error:
    - Manifested in $\mathbb{E}_\alpha[\mathbb{V}_\xi[f]]$ – Not reducible