SAND2012-0693C

Data Free Inference in Computational Models

Habib N. Najm

hnnajm@sandia.gov

Combustion Research Facility
Sandia National Laboratories, Livermore, CA

First Annual CESM Uncertainty Quantification and Analysis Interest Group Meeting
Boulder, CO; Jan 30-31, 2012

Acknowledgement

B.J. Debusschere, R.D. Berry, K. Sargsyan, C. Safta
— Sandia National Laboratories, CA
R.G. Ghanem — U. South. California, Los Angeles, CA
O.M. Knio — Duke Univ., Durham, NC
O.P. Le Maître — CNRS, Paris, France
Y.M. Marzouk — Mass. Inst. of Tech., Cambridge, MA

This work was supported by:

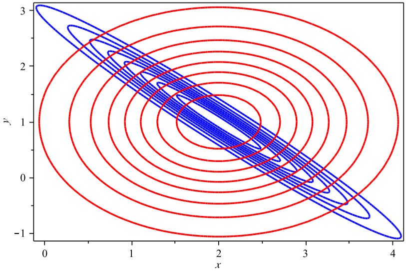
- US Department of Energy (DOE), Office of Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC)
- DOE Office of Basic Energy Sciences (BES) Division of Chemical Sciences, Geosciences, and Biosciences.
- DOE ASCR Applied Mathematics program.
- 2009 American Recovery and Reinvesment Act.

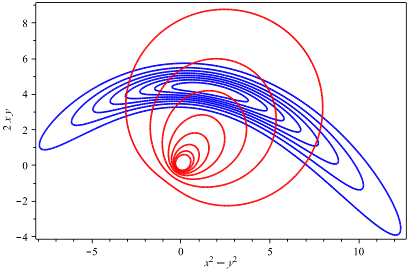
Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

Motivation

- Probabilistic UQ requires specification of uncertain inputs
- Require joint PDF on input space
- PDF can be found given data
- Typically such PDFs are not available from the literature
  - Summary information, e.g. nominals and bounds, is usually available
- Uncertainty in computational predictions can depend strongly on detailed structure of the missing parametric PDF
- Need a procedure to reconstruct a PDF consistent with available information in the absence of the raw data
  - “Data Free" Inference (DFI) (Berry *et al.*, JCP 2012)

The strong role of detailed input PDF structure





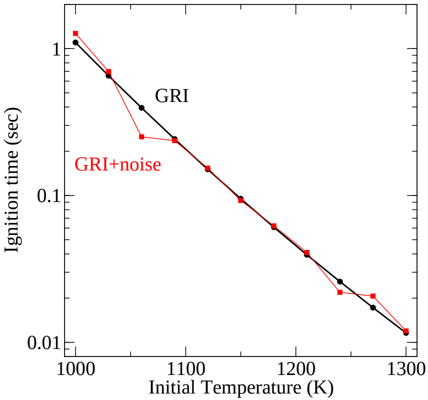
- Simple nonlinear algebraic model $(u, v) = (x^2 - y^2, 2xy)$
- Two input PDFs, $p(x, y)$
  - same nominals/bounds
  - different correlation structure
- Drastically different output PDFs
  - different nominals and bounds

Outline

Generate ignition "data" using a detailed model+noise

- Ignition using a detailed chemical model for methane-air chemistry
- Ignition time versus Initial Temperature
- Multiplicative noise error model
- 11 data points:

$$\begin{aligned} d_i &= t_{\mathrm{ig},i}^{\mathsf{GRI}}(1+\sigma\epsilon_i) \\ \epsilon &\sim N(0,1) \end{aligned}$$



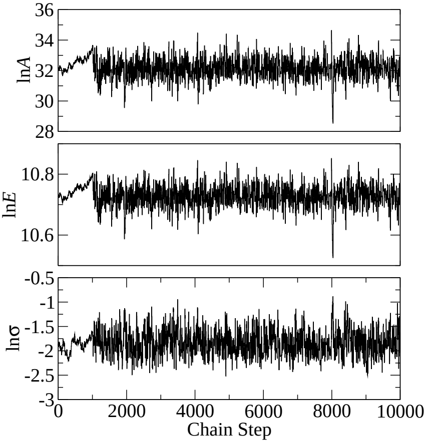
Fitting with a simple chemical model

- Fit a global single-step irreversible chemical model

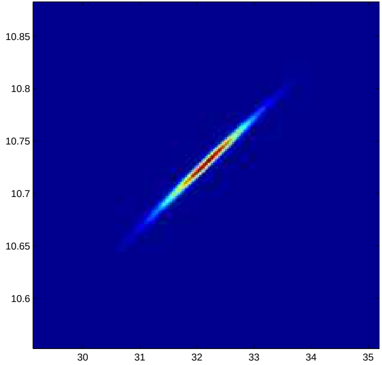
$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$$

$$\begin{aligned} \mathfrak{R} &= [CH_4][O_2]k_f \\ k_f &= A\exp(-E/R^oT) \end{aligned}$$

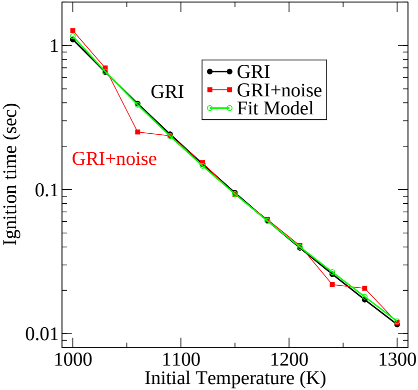
- Infer 3-D parameter vector $(\ln A, \ln E, \ln\sigma)$
- Good mixing with adaptive MCMC when start at MLE



Bayesian Inference Posterior and Nominal Prediction



Marginal joint posterior on $(\ln A, \ln E)$ exhibits strong correlation



Nominal fit model is consistent with the true model

Data Free Inference (DFI)

(Berry *et al.*, JCP 2012)

- Intuition: In the absence of data, the structure of the fit model, combined with the nominals and bounds, implicitly inform the correlation between the parameters
- Goal: Make this information *explicit* in the joint PDF
- DFI: discover a consensus joint PDF on the parameters consistent with given information:
  - Nominal parameter values
  - Bounds
  - The fit model
  - The data range
  - ... potentially other/different constraints

Data Free Inference Challenge

Discarding initial data, reconstruct marginal $(\ln A, \ln E)$ posterior using the following information

- Form of fit model
- Range of initial temperature
- Nominal fit parameter values of $\ln A$ and $\ln E$
- Marginal 5% and 95% quantiles on $\ln A$ and $\ln E$

Further, for now, presume

- Multiplicative Gaussian errors
- $N = 8$ data points

DFI Algorithm Structure

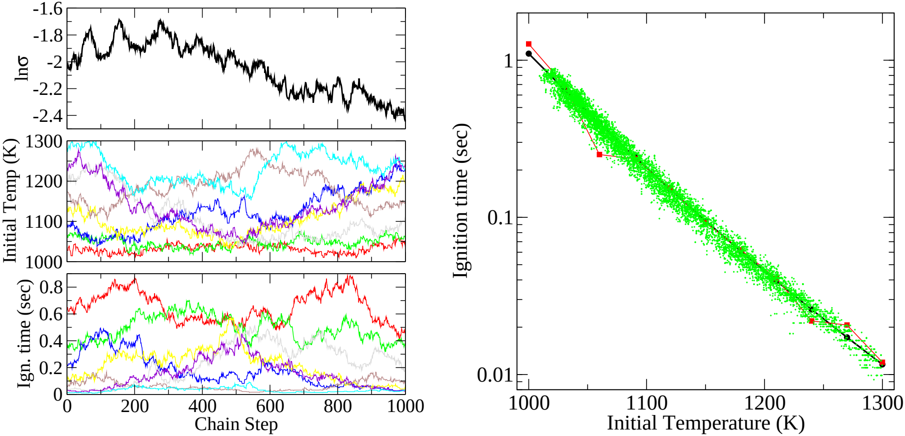
Basic idea:

- Explore the space of hypothetical data sets
  - MCMC chain on the data
  - Each state defines a data set
- For each data set:
  - MCMC chain on the parameters
  - Evaluate statistics on resulting posterior
  - Accept data set if posterior is consistent with given information
- Evaluate pooled posterior from all acceptable posteriors

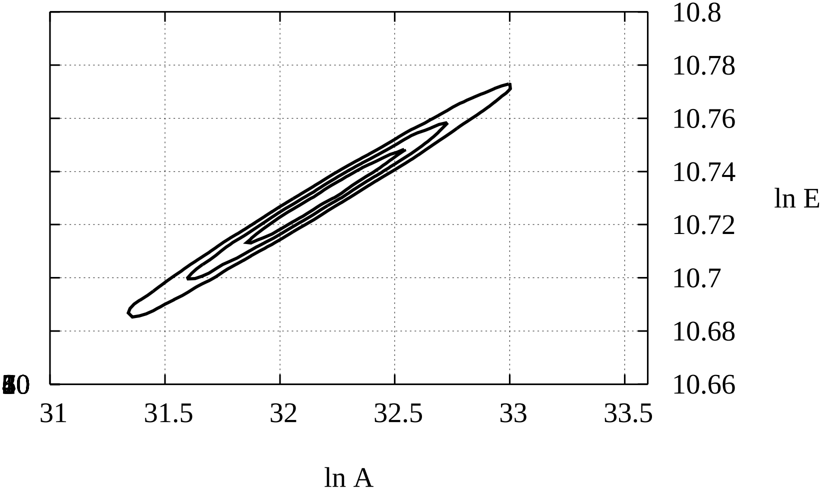
  Logarithmic pooling:

$$p(\lambda|y) = \left[\prod_{i=1}^{K} p(\lambda|y_i)\right]^{1/K}$$

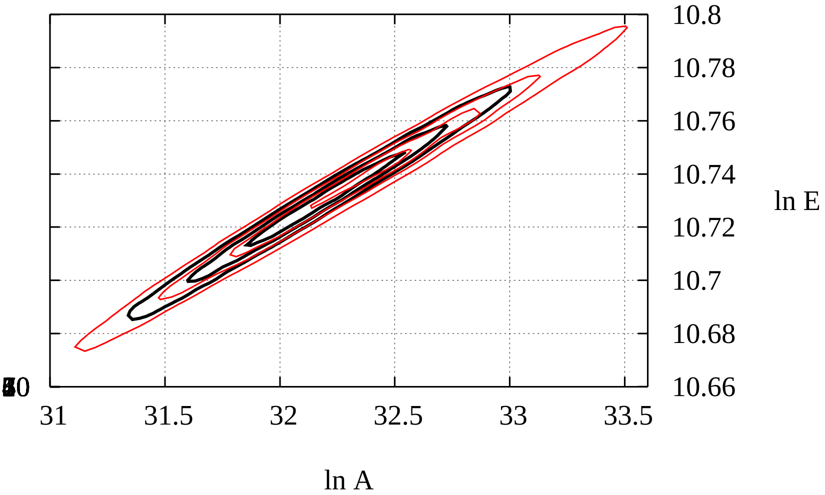
Short sample from outer/data chain



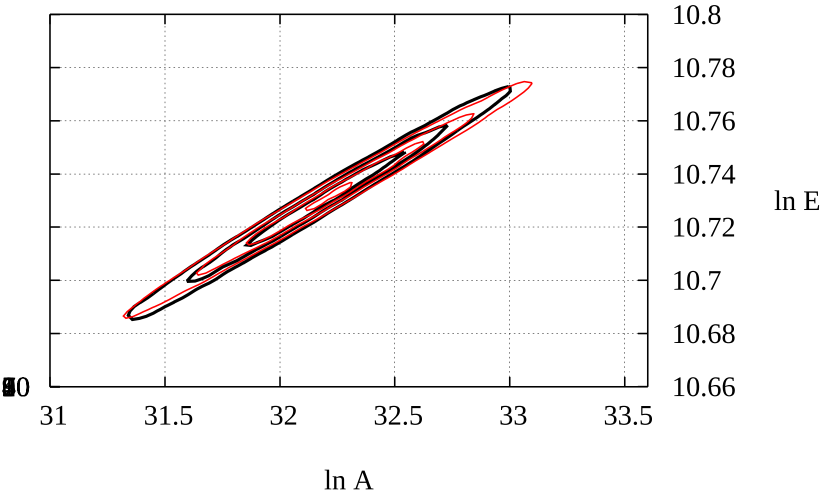
Reference Posterior – based on actual data



Ref + DFI posterior based on a 1000-long data chain



Ref + DFI posterior based on a 5000-long data chain



Closure

- Need for probabilistic characterization of uncertain inputs of climate models
  - Correlations important for uncertainty in predictions
- Given either old or new data
  - Bayesian inference can be used to provide the joint posterior PDF on model parameters
- In the absence of data
  - DFI $\Rightarrow$ joint PDF consistent with available information
    - Relationship to the Bayesian missing data problem, and maximum entropy estimation
  - Require information on experiments/instruments/fitting used to measure each parameter