

SAND2013-10309C

*Inference of Chemical Model Parameters given  
Partial Information*

H. Najm, R. Berry,  
K. Chowdhary, C. Safta, K. Sargsyan, & B. Debusschere

**[hnnajm@sandia.gov](mailto:hnnajm@sandia.gov)**

Sandia National Laboratories,  
Livermore, CA

MaxEnt2013  
Dec 15-20, 2013  
Canberra, Australia

# Acknowledgement

This work was supported by:

- DOE Office of Basic Energy Sciences, Div. of Chem. Sci., Geosci., & Biosci.
- DOE ASCR Applied Mathematics program.
- 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

- 1 Introduction
- 2 Data Free Inference
- 3 Processed Data Products
- 4 Closure

# Uncertainty in Physical Model Inputs

- Probabilistic UQ requires specification of uncertain inputs
- Require joint PDF on input space
- PDF can be found given data
  - Data is rarely published
- Typically such PDFs are not available from the literature
  - What is typically available are summaries, *e.g.* nominals and bounds on data, parameters, or fitted model
  - Processed data products
- Uncertainty in computational predictions can depend strongly on detailed structure of the parametric PDF
- Need a procedure to reconstruct a PDF consistent with available information in the absence of the raw data
  - Published summaries
  - Fitting model
  - Other known experimental details

# Bayes formula for Parameter Inference

- Data Model (fit model + noise):  $y = f(\lambda) + \epsilon$
- Bayes Formula:

$$p(\lambda, y) = p(\lambda|y)p(y) = p(y|\lambda)p(\lambda)$$

$$\begin{array}{c}
 \text{Likelihood} \quad \text{Prior} \\
 p(\lambda|y) = \frac{p(y|\lambda) p(\lambda)}{p(y)} \\
 \text{Posterior} \qquad \qquad \qquad \text{Evidence}
 \end{array}$$

- Prior: knowledge of  $\lambda$  prior to data
- Likelihood: forward model and measurement noise
- Posterior: combines information from prior and data
- Evidence: normalizing constant for present context

# Inference with Partial/Missing Data – Scenarios

- Partial data: some of the data is missing
  - *e.g.* surveys, census, *etc*
  - ⇒ Imputation methods
- All the data is missing but some information is known
  - Moments/statistics, or summaries
    - on missing data
    - on parameters fitted to missing data
    - on outputs of models fitted to missing data
  - Physical constraints
  - Details about the experiment
  - ⇒ Maximum Entropy principle

# No Data – Inference based on Constraints

Maximum Entropy (MaxEnt) Principle:

Maximize ignorance while satisfying given constraints  $\mathcal{C}$

Given data  $z$ , use Bayes rule to get

$$q(\beta|z) = \frac{q(z|\beta)q(\beta)}{q(z)}$$

When data  $z$  is unavailable, but its distribution  $w(z|\mathcal{C})$  can be estimated, then, given a prior  $q(z, \beta)$ , maximizing relative entropy  $\mathcal{E}(p, q)$  leads to the joint MaxEnt posterior

$$p(z, \beta|\mathcal{C}) = q(\beta|z)w(z|\mathcal{C})$$

Marginalizing over  $z$ :  $p(\beta|\mathcal{C}) = \int q(\beta|z)w(z|\mathcal{C})dz$

# Data Free Inference (DFI)

(Berry *et al.*, JCP 2012)

- **Intuition:** In the absence of data,
  - the structure of the fit model
  - nominals and bounds
  - physical constraints
  - other known experimental detailsimplicitly 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
  - Maximum entropy principle
  - Approximate Bayesian Computation (ABC) methods

# DFI Algorithm in General

Given fitted model  $f(x, \beta)$  and statistics from missing data  $z^*$   
 $z^* \Rightarrow p(\beta|z^*) \Rightarrow \mathcal{S}(G(p(\beta|z^*))) \Rightarrow \mathcal{S}(z^*) := \mathcal{S}^*$

- MCMC chain on data space
  - Each chain step  $\Rightarrow$  data  $z$
  - Likelihood  $\pi_\epsilon(z) = K_\epsilon(|\mathcal{S}(z) - \mathcal{S}^*|)$ 
    - $\epsilon \rightarrow 0 \Rightarrow \{\pi_\epsilon(z) \neq 0 \Leftrightarrow \mathcal{S}(z) \rightarrow \mathcal{S}^*\}$
    - $\mathcal{S}(z)$  evaluated from the inner MCMC chain on the parameter space
- MCMC chain on parameter space
  - Estimate posterior  $p(\beta|z)$
  - Evaluate statistic  $\mathcal{S}(z)$
- Pooling/marginalizing consistent posteriors  $p(\beta|z)$

# DFI Demonstration – Plan

Context: Chemical ignition in a homogeneous CH<sub>4</sub>-Air system

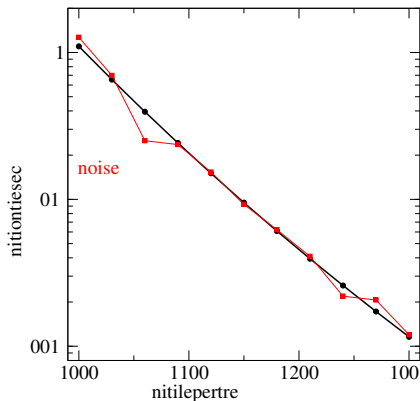
- Use simulated noisy data to infer a joint posterior PDF on parameters of a chemical model
- Discard data, retaining summary information
- Reconstruct the posterior using available information
- Compare the “true” and DFI posteriors

# 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:

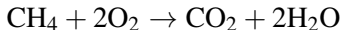
$$\tau_i^d = \tau^{\text{GRI}}(T_i^o) (1 + \sigma \epsilon_i)$$

$$\epsilon \sim N(0, 1)$$



# Fitting with a simple chemical model

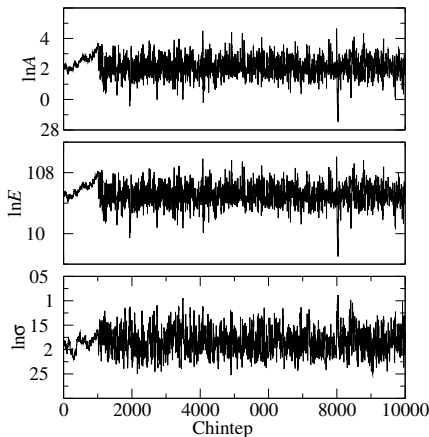
- Fit a global single-step irreversible chemical model



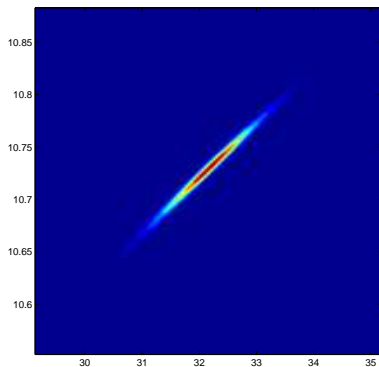
$$\mathfrak{R} = [\text{CH}_4][\text{O}_2]k_f$$

$$k_f = A \exp(-E/R^oT)$$

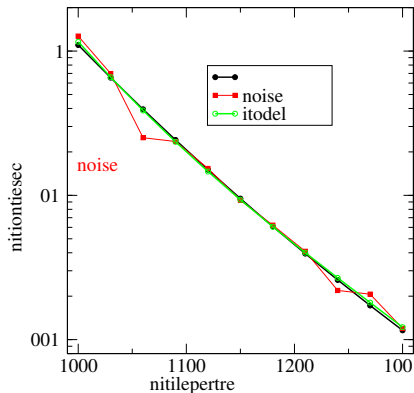
- Infer 3-D parameter vector  $(\ln A, \ln E, \ln \sigma)$
- Good mixing with adaptive MCMC when start at MLE



# 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 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

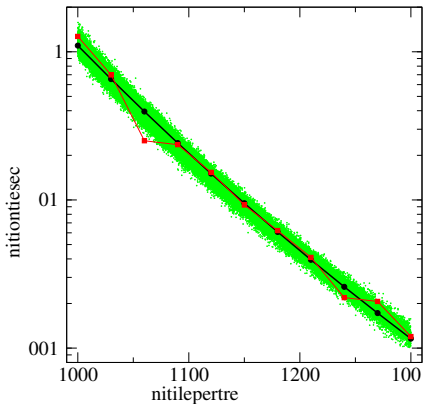
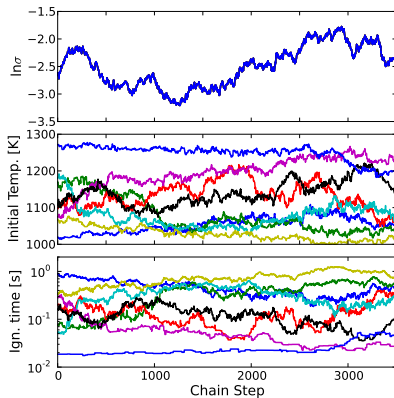
- Explore the space of hypothetical data sets
  - MCMC chain on the data
  - Each state defines a data set
- For each data set:
  - MCMC chain on the parameters
  - Evaluate statistics on resulting posterior
  - Accept data set if posterior is consistent with given information
- Evaluate pooled posterior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

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

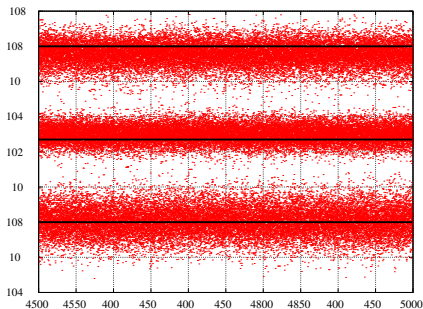
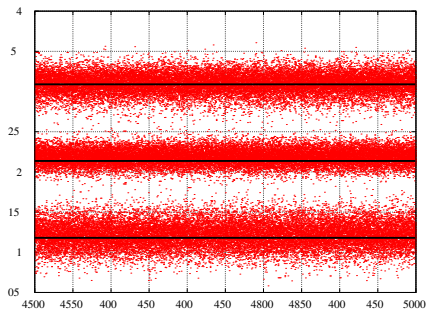
# Computational issues

- Outer Chain:
  - $N$  data points  $\Rightarrow$  chain is  $(2N + 1)$ -dimensional
    - Single-site update – easy to tune in hi-D
  - Markov chain on a manifold in data space
    - Ideally, use geometric MCMC methods
    - No sense of good mixing – no peak
    - Pursue convergence with # of samples
- Inner Chain:
  - Highly peaked posterior ridge
    - Adaptive MCMC for parameter estimation
    - Start at nominal model
    - Ideal proposal distrib. uses Hessian of Likelihood
- Computationally challenging:
  - Use a Polynomial Chaos surrogate for forward model
  - Run multiple outer chains in parallel, and aggregate resulting acceptable data sets

## Short sample from outer/data chain



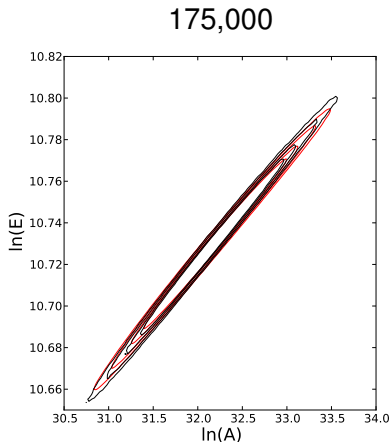
# Scatter of Nominals and 0.05/0.95 Quantiles



- Parameter posterior statistics are in general agreement with the desired nominal/quantile summaries
- Amplitude of scatter can be controlled by choice of weights in the likelihood function of the data chain

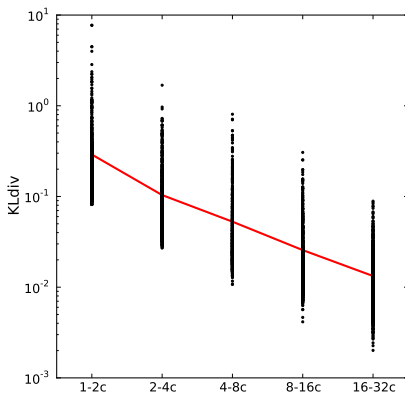
# Pooled DFI Posterior vs. Reference Posterior

- 50 data chains run in parallel  
⇒ Pooled DFI posterior
- Structure of 2D pooled posterior is robust
  - Irrespective of the number of chains
- Very close similarity to the reference posterior
- Available information is essentially a sufficient statistic



# Empirical Convergence of 3D Pooled Posteriors

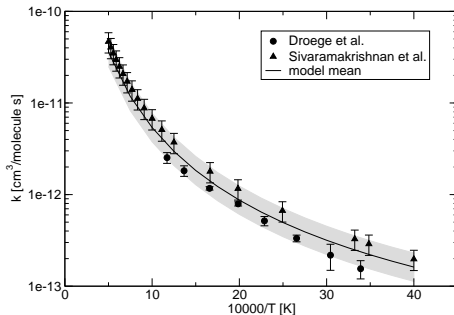
- Kullback-Leibler divergence between posteriors of successively increased data volumes
- Combinatorial choices of chains pooled at each stage
  - statistical scatter of KLdiv
- Overall convergence evident  $\propto 1/N$



# Processed Data Products

Frequently, we are given published “data” in the form of nominal values and error bars

- Fitting of  $y = f(x; \beta)$
- Data:  
 $(x_\ell, y_\ell^o \pm \alpha_\ell)$ ,  $\ell = 1, \dots, L$   
 $\pm \alpha_\ell$  are e.g.  $\pm 2\sigma$  error bars



Two interpretations of the  $(y_\ell^o, \alpha_\ell)$  as nominals and bounds

- on the **data**:  $N$  observed random  $y_\ell$  at each  $x_\ell$
- on the **fitted model**: uncertain  $y_\ell = f(x_\ell, \beta)$

# Processed Data Products – Demo

Demonstrate DFI on a simple model problem:

- Generate synthetic noisy data from a double exponential decay model

$$s = 2e^{-2t} + .5e^{-.5t}$$
$$y = s + (\sigma_1 s + \sigma_2)\eta$$

Retain only error bars at different times

- Fit using DFI with a single exponential decay model

$$f = a \exp(-kt)$$
$$y = f + (\sigma_1 f + \sigma_2)\eta$$

# DFI – Error Bars on Model Outputs

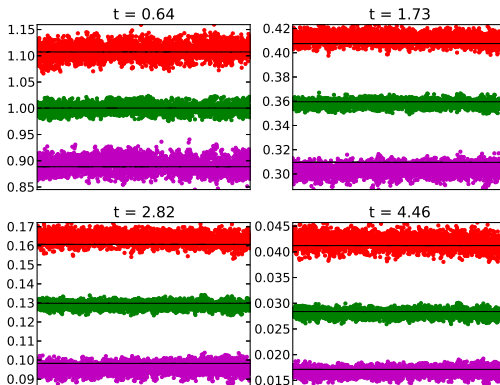
Given required statistics  $\mathcal{S}^*$

Missing data  $z^*$

$$z^* \Rightarrow p(\beta|z^*) \Rightarrow p(f(x, \beta)|z^*) \Rightarrow \mathcal{S}(z^*) := \mathcal{S}^*$$

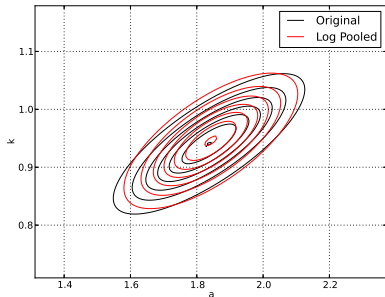
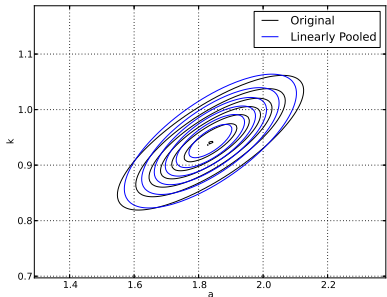
- MCMC chain on data space
  - Each chain step  $\Rightarrow$  data  $z$
  - Likelihood  $\pi_\delta(z) = f_\delta(\mathcal{S}(z) - \mathcal{S}^*)$ 
    - $\mathcal{S}(z)$  evaluated from the inner MCMC chain on the parameter space
- MCMC chain on parameter space
  - Estimate posterior  $p(\beta|z)$
  - Estimate pushed forward posterior  $p(f(x, \beta)|z)$
  - Evaluate statistic  $\mathcal{S}(z)$
- Pooling

# PFP Quantiles



- Consistent pushed forward posterior quantiles clustered in the vicinity of reference values

# Linear and Logarithmic Pooling give Similar Results



- Posteriors from individual data sets are largely similar to the reference posterior, hence the similar pooling results

# Closure

Outlined several missing data contexts

- Partial data – Imputation
- No data
  - Maximum Entropy principle
  - Approximate Bayesian Computation
  - DFI
- Demonstrations
  - Constraints on fitted parameters of interest
  - Processed data products
    - Constraints on raw data
    - Constraints on fitted model uncertain predictions