This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2018-11882C

Uncertainty Quantification with Missing Data

Habib N. Najm

Sandia National Laboratories
Livermore, CA, USA
hnnajm@sandia.gov

Seminar

Sapienza University of Rome
Rome, Italy
Oct 31, 2018

andia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly own subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Acknowledgement

B.J. Debusschere, M. Reagan, R.D. Berry, K. Sargsyan, C. Safta,
K. Chowdhary, M. Khalil, X. Huan, M. Eldred, G. Geraci, T. Casey, J. Oefelein,
G. Lacaze, Z. Vane, L. Hakim
– Sandia National Laboratories, CA

R.G. Ghanem – U. South. California, Los Angeles, CA
O.M. Knio – KAUST, Thuwal, Saudi Arabia & 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:

- DOE Office of Basic Energy Sciences, Div. of Chem. Sci., Geosci., & Biosci.
- DOE Office of Advanced Scientific Computing Research (ASCR)
- DOE ASCR Scientific Discovery through Advanced Computing (SciDAC) program
- DOE ASCR Applied Mathematics program
- DARPA
- Sandia National Laboratories, LDRD

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

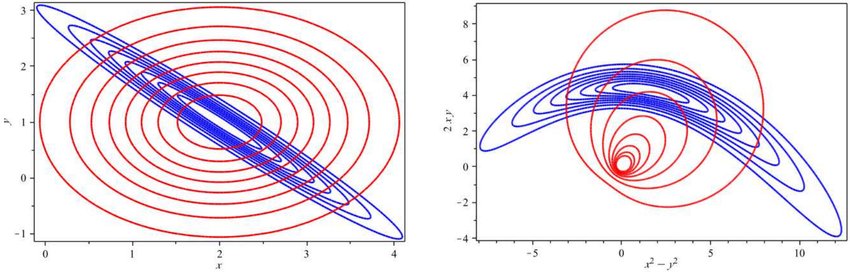
Outline

1. Introduction
2. Algorithm
3. Application
4. Closure

Uncertainty in Model Inputs

- 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)

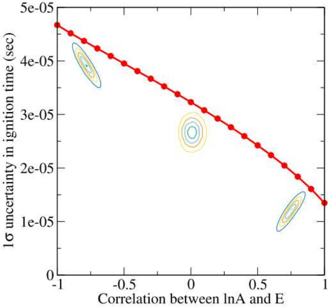
The strong role of detailed input PDF structure



- 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

Correlations among uncertain Arrhenius rate parameters are important for accurate prediction with uncertainty

- Uncertainty quantification in combustion requires estimation of the *joint* uncertainty structure on model parameters
  - most importantly - rate constants - $k(T) = AT^n \exp(-E/RT)$
- Published kinetics literature typically includes
  - Error bars on ln A, sometimes on E – no information on correlations
- Ignoring the correlation among (A, n, E) leads to over/under prediction of uncertainty in combustion model outputs
- Strong dependence of standard deviation in predicted ignition time for hydrogen on the correlation between the uncertain $(\ln A, E)$ of the rxn: $H + O_2 \longrightarrow OH + O$
  - holding marginal $\sigma_{\ln A}, \sigma_E$ constant



Dealing with Missing Data – MaxEnt

Maximum Entropy Principle:

Maximize uncertainty while satisfying given constraints

Bayes rule

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

When data z is unavailable,
but its distribution, conditioned on constraints S, $w(z|S)$ is known, then,
given a prior $q(z, \lambda)$,
maximizing relative entropy $\mathcal{E}(p, q)$ leads to the joint MaxEnt posterior

$$p(z, \lambda|S) = q(\lambda|z)w(z|S)$$

Marginalizing over z, arrives at the marginal parameter posterior $p(\lambda|S)$

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

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 statistics
- Evaluate pooled posterior from all acceptable posteriors
  - Logarithmic pooling:

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

  - Linear pooling

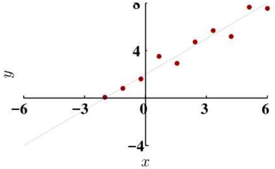
$$p(\lambda|z) = \frac{1}{N}\sum_{i=1}^{K} p(\lambda|z_i)$$

DFI Uses two nested MCMC chains

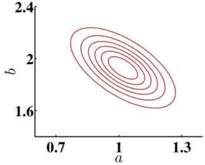
- An outer chain on the data, $(2N+1)$–dimensional
  - Generally high-dimensional
  - N data points (x_i, y_i) + σ
  - Likelihood function captures constraints on parameter nominals+bounds
- An inner chain on the model parameters
  - Conventional MCMC for parameter estimation
  - Likelihood based on fit-model
  - parameter vector $(\ln A, \ln E, \ln \sigma)$
- Computationally challenging
  - Single-site update on outer chain
  - Adaptive MCMC on inner chain
  - Run multiple outer chains in parallel, and aggregate resulting acceptable data sets

DFI Illustration: Linear Regression – Problem setup

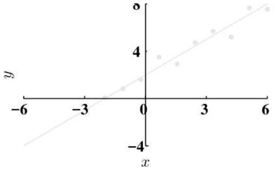
- Using Bayesian inference, one can estimate parameters given data



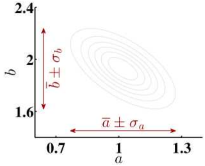
$$y_i = ax_i + b + \epsilon_i$$



- Here, however, neither data nor the joint parameter posterior is reported
- Available published "data" is in the form of 1-D summary statistics
  - *e.g.* mean, standard deviation



$$y_i = ax_i + b + \epsilon_i$$

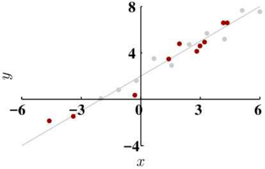


DFI Illustration: Linear Regression – Algorithm

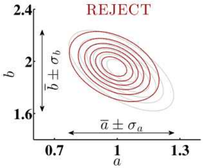
Basic idea:

- Infer both data *and* parameters that satisfy given constraints ⇒ $p(\lambda, z|S)$
- Average/marginalize over data space ⇒ $p(\lambda|S)$

- Sample over data space using MCMC ⇒ z^i, $i = 1, 2, ...$
- Estimate Bayesian posterior $p(\lambda|z^i)$ on parameters given data set sample
- Accept/reject data sample based on how well the parameter posterior matches the given summary statistics S



$$y_i = ax_i + b + \epsilon_i$$



- Marginalize over z – pooling of consistent posteriors ⇒ $p(\lambda|S)$

Linear Regression Example

- Inference of parameter a provided μ_a and σ_a

$$y_i = ax_i + b + \epsilon_i \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

- The posterior parameter moments relate to the data through

$$\mu_a = \sigma^2 \frac{\sum x_i y_i}{\sum x_i^2} \qquad \sigma_a = \sqrt{\frac{\sigma^2}{\sum x_i^2}}$$

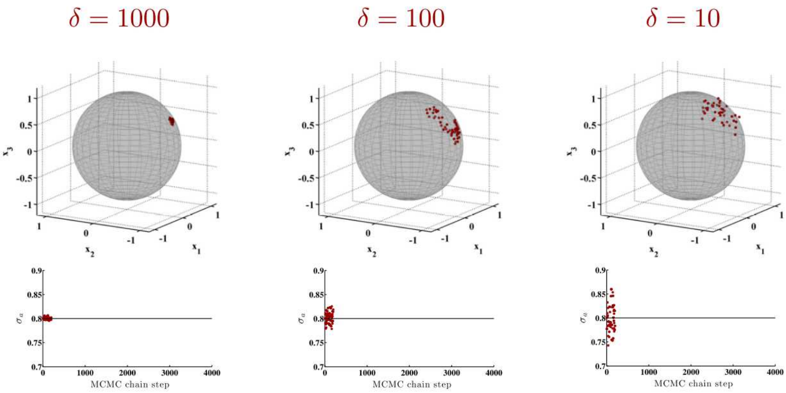
- In the absence of data $\{x_i, y_i\}, i = 1, \ldots, N$, assume $N = 3$ data points were used in the original calibration exercise. The standard deviation constraint results in

$$x_1^2 + x_2^2 + x_3^2 = \frac{\sigma^2}{\sigma_a^2}$$

MCMC Sampling with ABC Likelihood

- We relax the moment constraint using an ABC data likelihood

$$
\begin{aligned}
p_{\text{abc}}(x_1, x_2, x_3 | F) &\propto \exp\left[-\delta\left(F_i - F\right)^2\right] \\
&= \exp\left[-\delta\left(\sqrt{\sigma^2/\left(x_1^2 + x_2^2 + x_3^2\right)} - \sigma_a\right)^2\right]
\end{aligned}
$$



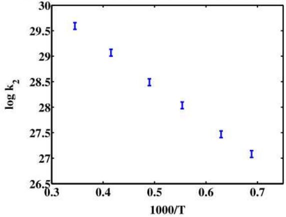
$\delta = 1000$ $\delta = 100$ $\delta = 10$

Shock tube study – available statistics

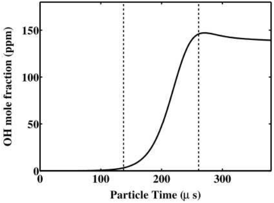
- Masten et al. (1990) performed shock tube experiments to measure reaction rate of reaction $H + O_2 \longrightarrow OH + O$
- Rate coefficient k determined by fitting the rapid OH growth region

T, K	P, atm	H_2 mol %	O_2 mol %	$\log k$	standard deviation
1452	1.461	4.99	0.500	27.08	0.022
1589	0.606	4.97	0.496	27.47	0.022
⋮	⋮	⋮	⋮	⋮	⋮
2769	1.991	0.964	0.200	29.59	0.022

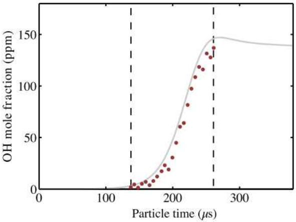
Published rate statistics



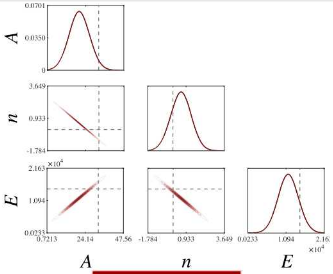
Unpublished species profile



Algorithm application to shock tube data

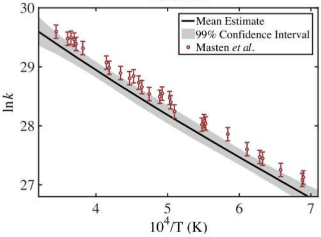


Bayesian inference



Propagate uncertainty

REJECT



- MCMC sample over the data space
  - OH concentration profiles
- Given data, infer joint parameter posterior
- Evaluate associated statistics on $k(T)$
- Accept/reject based on match between published/generated statistics
- Allow for uncertainty in pre-exponential of $OH + H_2 \longrightarrow H_2O + H$
- Pool all consistent posteriors

Numerous Technical Challenges

- High dimensional data space, many OH data points over a number of operating conditions
  - Optimization strategies to approach consistent data manifold
- Expensive forward model
  - **TChem** thermochemistry library adapted for fast sampling
    http://www.sandia.gov/tchem
  - **UQTk** UQ library for efficient and flexible PC UQ
    www.sandia.gov/UQToolkit
  - Padé-Polynomial Chaos model surrogates
  - Quadrature evaluation of parameter posterior moments
- Pre-exponential nuisance parameter
  - Formulation with consistent estimation of model evidence
    Khalil & Najm, CTM, submitted
- Serial nature of MCMC methods
  - Many independent data chains in parallel

Computation of Expectations

- For each data set z_i, to compute the ABC likelihood, need the expectation

$$F_i = \int d\lambda \, p\left(\lambda | z_i\right) f\left(\lambda\right)$$

- Can be done via MC integration with MCMC samples of λ from $p\left(\lambda | z_i\right)$
- However, in this case, despite the strong nonlinearity of the model, all consistent data sets result in nearly Gaussian posterior parameter pdfs
- Use importance sampling with a Gaussian proposal q

$$F_i \quad = \int d\lambda \, \frac{p(\lambda|z_i)}{q(\lambda|z_i)} f\left(\lambda\right) q\left(\lambda|z_i\right)$$

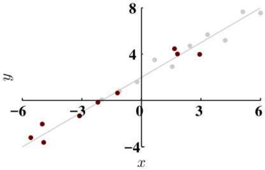
- Approximate the above integral using Gauss-Hermite quadrature

$$F_i \quad = \sum_{j=1}^{N_q} w_j \frac{p(\lambda_j|z_i)}{q(\lambda_j|z_i)} f\left(\lambda_j\right)$$

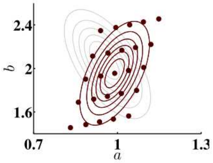
- Number of evaluations of the parameter posterior p, involving forward model simulations, are reduced by 3-4 orders of magnitude

Adaptive Gauss-Hermite Quadrature

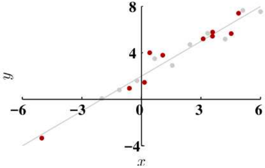
- Assuming that we have the posterior mean and covariance matrix of the parameter vector for an initial data (via one initial MCMC simulation)
- Compute a quadrature rule which will be used for importance sampling



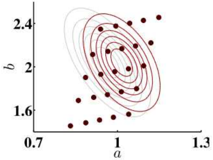
$$y_i = az_i + b + \epsilon_i$$



- For each proposed data set, iteratively compute the posterior mean and covariance matrix using the available quadrature rule, and recompute the quadrature rule with the updated mean and covariance



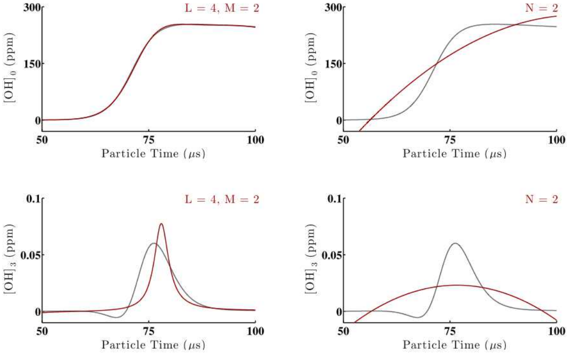
$$y_i = az_i + b + \epsilon_i$$



Use of hybrid Padé/Polynomial Chaos surrogates

- Model $[OH](k,t)$ using Pol. Chaos in parameter space; Padé in time
  - exhibiting strong nonlinearity
- Domain-decomposition in (k_1,k_2) space for minimal RMSE, given order

$$[\mathrm{OH}](k_1,k_2,t) \simeq \sum_{i=0}^{N} [\mathrm{OH}]_i(t)\Psi_i(k_1,k_2) = \sum_{i=0}^{N} \frac{\sum_{m=0}^{M} p_{m,i}\Phi_m(t)}{\sum_{l=0}^{L} q_{l,i}\Phi_l(t)}\Psi_i(k_1,k_2)$$



Padé PCE

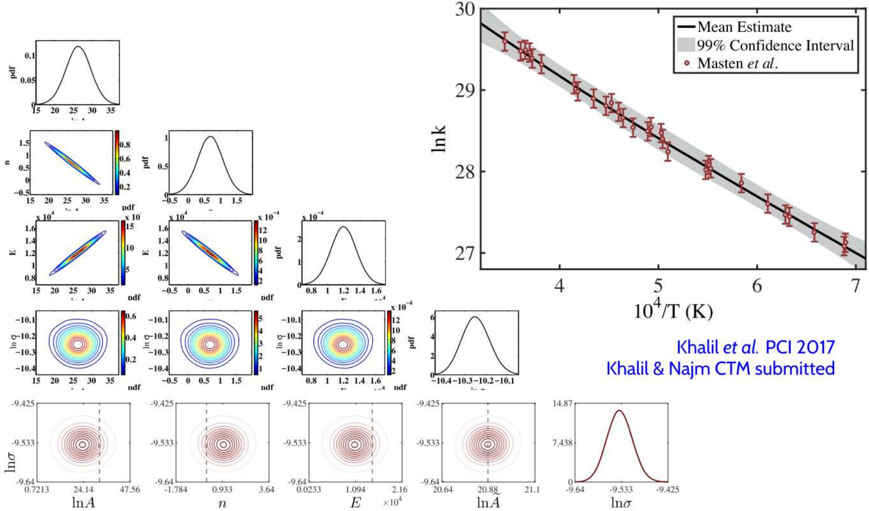
Nuisance Parameters with Prescribed PDF

- The experimentalists performed the calibration without infering nuisance parameters (rate coefficient of OH + H2 → H2O + H).
- They fixed the level of uncertainty in these parameters while performing the calibration.
- Challenge: Infer the joint probability destribution of all parameters with nuisance parameters having prescribed PDF $\pi\left(\phi\right)$

$$\begin{aligned} p\left(\lambda, \phi \,|\, z\right) &= p\left(\lambda \,|\, z, \phi\right) p\left(\phi \,|\, z\right) \\ &= p\left(\lambda \,|\, z, \phi\right) \pi\left(\phi\right) \\ &= \frac{p\left(z \,|\, \lambda, \phi\right) p\left(\lambda \,|\, \phi\right)}{p\left(z \,|\, \phi\right)} \pi\left(\phi\right) \end{aligned}$$

  - Discretize ϕ-space
  - Evaluate $p(\lambda, \phi_j \,|\, z), j = 1, \ldots J$
  - Interpolate for arbitrary ϕ

DFI result: pooled posterior parameter PDF



Khalil *et al.* PCI 2017

Khalil & Najm CTM submitted

Extensions to Multiple Models & Experiments

- Going beyond a single reaction/experiment
  - Rxn $r \Rightarrow p(\lambda_r|S_r, M_{\mathsf{e}}^r)$
  - Rxn $s \Rightarrow p(\lambda_s|S_s, M_{\mathsf{e}}^s)$
  - The experimental models may include reactions that are not present in model M employed for prediction
  - They may use different values of the same nuisance parameters
- The fundamental kernel for each DFI-inverted experiment is the set of consistent data-sets (z^r, z^s)
- Employ model M to pool consistent data sets from all experiments
  - Presuming it is valid for the experimental conditions
- Can have experiments at different conditions for estimating a given parameter
  - Employ Bayesian inference with multiple data sets
    - within the pooling framework

Algorithm Stage I – Sampling of Consistent Data Sets

Data: z
Parameters: λ
Experiment measures data z_{e} and fits it using model M_{e}
Constraints: Reported statistics $S^* = S(p(\lambda|z_{\mathsf{e}}, M_{\mathsf{e}}))$

Consistent data sets:

$$z \sim p(z|S^*, M_{\mathsf{e}})$$

Sample this density using MCMC on the data space, using ABC with a kernel-density pseudo-likelihood

$$p_{\mathsf{ABC}}(z|S, M_{\mathsf{e}}) \propto \mathsf{exp}(-\delta\|S(z; M_{\mathsf{e}}) - S^*\|_2^2)$$

where

$$S(z; M_{\mathsf{e}}) := f(p(\lambda|z, M_{\mathsf{e}}))$$

⇒ accumulate consistent data sets $(z_1, \dots, z_N)$, being the samples in the data MCMC chain

Algorithm Stage II – Pooling of Consistent Posteriors

Pooling operator T

$$p = T(p_1, \dots, p_N)$$

where

$$p_i = p(\lambda | z_i, M)$$

Linear average pooling:

$$p = \frac{1}{N} \sum_{i=1}^{N} p_i$$

Logarithmic average pooling:

$$p = \left[\prod_{i=1}^{N} p_i \right]^{1/N}$$

Two distinct pooling/prediction scenarios

- Employ same model $M = M_\mathrm{e}$
- Employ some alternate model M, similarly valid for the experiment physical system

Dealing with Multiple –independent– Experiments

Consider K independent experiments, under potentially different conditions, informing λ

$$p(\lambda|z_{\mathsf{e}}^1,\dots,z_{\mathsf{e}}^K,M_{\mathsf{e}}^1,\dots,M_{\mathsf{e}}^K)=\prod_{k=1}^{K}p(\lambda|z_{\mathsf{e}}^k,M_{\mathsf{e}}^k)$$

The task of DFI then is to generate consistent N data sets $(z_i,\dots,z_N)$ where each z_i is an instance of consistent data sets across all K experiments

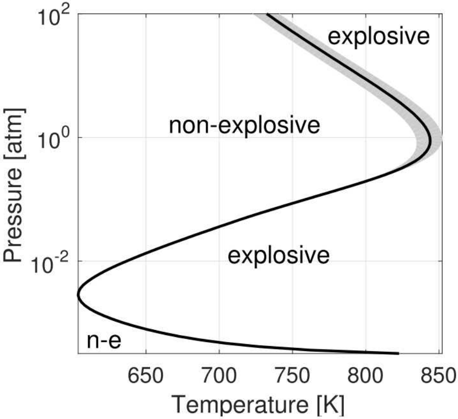
$$z_i=\{z_i^1,\dots,z_i^K\}$$

employing K reported statistics, associated ABC likelihoods, & data chains

The pooling operation employs consistent posteriors $(p_1,\dots,p_N)$:

$$p_i=\prod_{k=1}^{K}p(\lambda|z_i^k,M)$$

Application: H_2-O_2 explosion limits

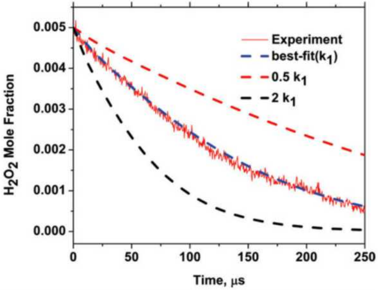


Stoichiometric H_2-O_2 explosion limit curve

- H_2O_2 decomposition controls the 3rd explosion limit in the H_2-O_2 system
- forward model: homogenous ignition delay
- binary classifier establishes the limit curve within 0.1 K.

H_2O_2 decomposition - target experiment

- $H_2O_2 + M \longrightarrow 2\,OH + M$
- Shock tube study, laser absorption by H_2O_2 [1]
- Rate constant determined by solving ODE system with a chemical mechanism [2]
- Data are concentration vs. time decay profiles of H_2O_2
- Reported fit: $k(T)=10^{16.29\pm0.12}\exp(-21993\pm301/T)$
- Also reported nominals & error bars on k(T) for a number of temperatures
  - Use these!



P [atm]	T [K]	$H_2O_{2\text{initial}}$ [%]
1.15	1012.5	0.42
1.08	1186.8	0.45
2.33	997.54	0.47
2.35	1166.6	0.50

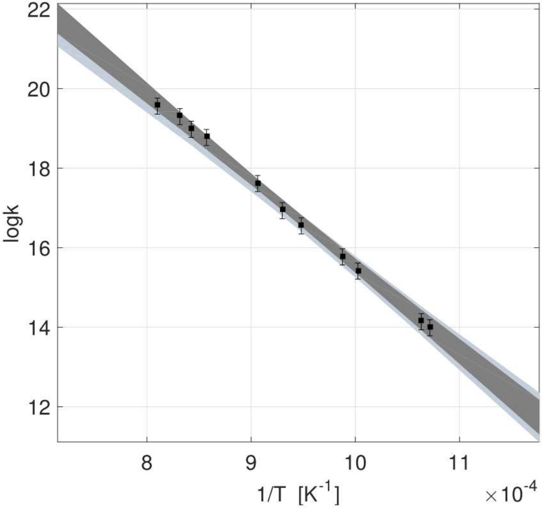
Table: Experimental conditions

[1] Sajid et al. Int. J. Chem. Kinet. (2014)
[2] Hong et al. Combustion and Flame 158 (2011)

Push-Forward Pooled Posterior on $k(T)$ for H_2O_2 + M Rxn

Low pressure data

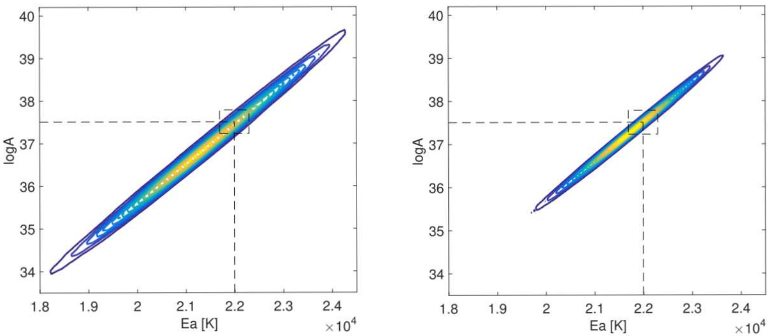
- Predictive distribution of the uncertain ln $k(T)$ overlaps reported nominal experimental measurements, and error bars



– Shading indicates one/two-σ interpretation –

Pooled Posterior density $p(\ln A, E)$ of H_2O_2 + M Rxn

Low pressure data



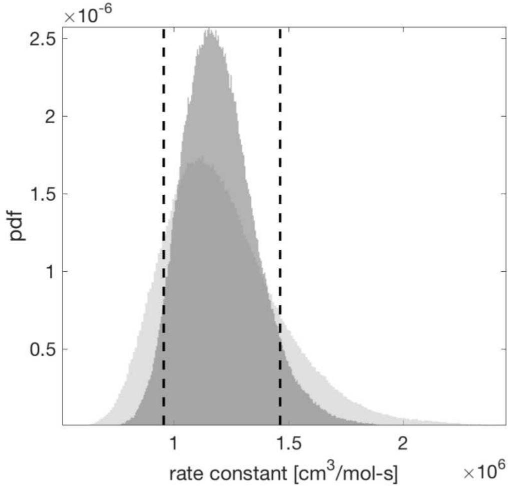
$k(T)$ error bars = 1σ

$k(T)$ error bars = 2σ

- MAP estimate close to reported nominal parameter values
- Posterior uncertainty much larger than that reported in quoted measurement of $(\ln A, E)$
- Found out subsequently that those reported $(\ln A, E)$ uncertainties were based on variability of the *nominal* $k(T)$ values

Pooled Posterior Density of $k(T)$ at $T = 933.3\ K$

Low Pressure



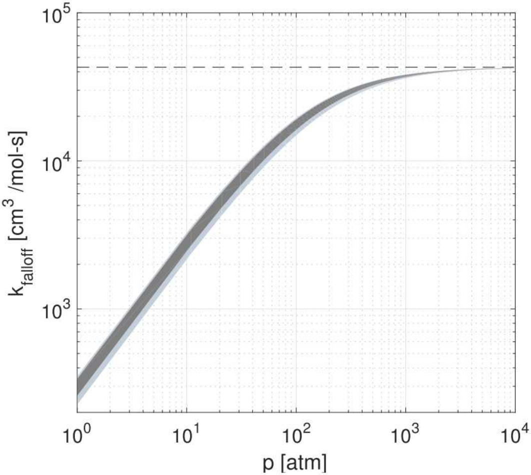
– Shading indicates one/two-σ interpretation –

– Dashed lines are exptl error bars –

Pooled Posterior of Rate – Dependence on Pressure

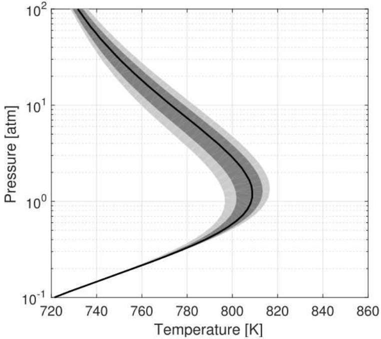
Low-pressure data

- The rate of this reaction in fact involves a dependence on pressure
- Augment existing low-P data with another experiment at higher pressure

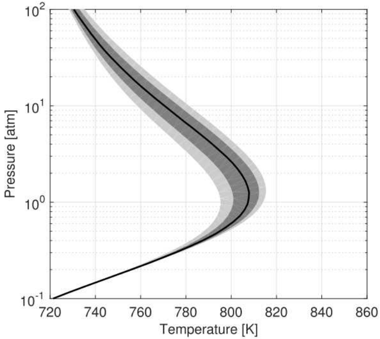


– Shading indicates one/two-σ interpretation –

Uncertainty in predicted explosion limit – Multiple data



Low P



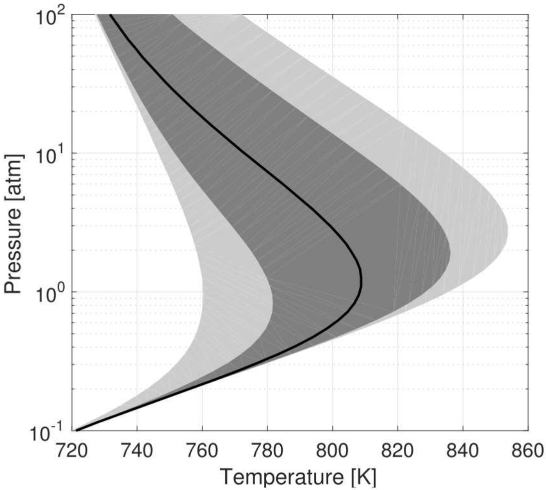
Low and intermediate P

– Shading indicates one/two-σ interpretation –

- Minimal impact of additional data in this case

Uncertainty in predicted explosion limit

- Presuming *iid* Gaussian inputs $p(\ln A), p(E)$
- Ignoring correlations among uncertain parameters can have a drastic impact on uncertainty in predictions



– Shading indicates one/two-σ interpretation –

Closure

- Need for parameter estimation in chemical systems given summary statistics
- DFI algorithm, based on MaxEnt and ABC methods
- Computationally challenging
- Extensions to multiple experiments and data sets
- On path towards full characterization of the parameters of H_2 oxidation based on experiments in the literature