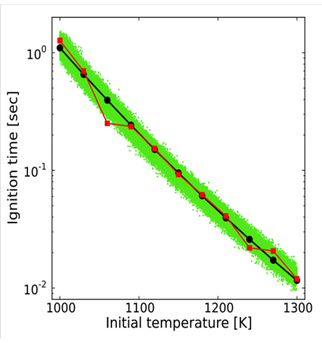
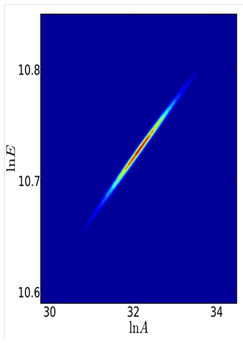
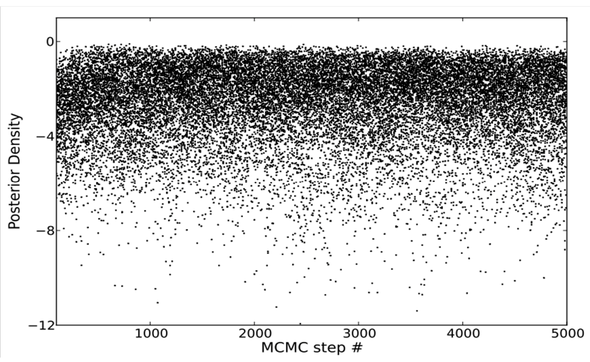






Sandia National Laboratories

Exceptional service in the national interest

Uncertainty Quantification

- A Process Based on Bayesian Inference

JK Shao, ERC, University of Wisconsin-Madison, MS. Mechanical Engineering, May 2014
Stanford University, Ph.D. Mechanical Engineering, est. Jun 2017
Mohammad Khalil, 8013, Combustion Research Facility, Sandia National Labs
Habib Najm, 8013, Combustion Research Facility, Sandia National Labs

SAND2014-15893D

Abstract: *This project focuses on uncertainty quantification based on Bayesian inference. A brief introduction of Bayesian inference is given in the first slide. In addition, one common method people usually use for uncertainty quantification is the least-square method. Thus, the comparison between least-square method and maximum likelihood function method is introduced in the second slide. A sampling method, Markov Chain Monte Carlo (MCMC), which is used for both the forward problem and the inverse problem, is then illustrated in the third slide. Based on Bayesian inference and MCMC sampling, a forward problem of uncertainty quantification is conducted. Parameter estimations based on some data of a turbulent flame experiment for the comparison of three different turbulent time scales are demonstrated in this forward problem. After that, an inverse problem is introduced. With Data Free Inference (DFI) code developed in CRF, an inverse problem of uncertain parameters in chemical models will be investigated in this summer.*

What is Bayesian inference?

Bayes' theorem and marginalization:

$$prob(X|Y,I) = \frac{prob(Y|X,I) \times prob(X|I)}{prob(Y|I)}$$

and

$$prob(X|I) = \int_{-\infty}^{+\infty} prob(X,Y|I)dY$$

The importance of this property to data analysis becomes apparent if X and Y are replaced by hypothesis and data:

$$prob(hypothesis|data,I) \propto prob(data|hypothesis,I) \times prob(hypothesis|I)$$

The power of Bayes's theorem lies in the fact that it relates the quantity of interest, the probability that the hypothesis is true given the data, to the term having a better chance of being able to assign, the probability that should be observed from the measured data if the hypothesis was true.

The various terms in Bayes' theorem have formal names. The quantity on the far right, prob(hypothesis|I), is called the prior probability; it represents our state of knowledge about the truth of the hypothesis before we have analyzed the current data. This is modified by the experimental measurements through the likelihood function, or prob(data|hypothesis,I), and yields the posterior probability, prob(hypothesis|data,I), representing our state of knowledge about the truth of the hypothesis in the light of the data. It should be note, however, that the equality of Bayes's theorem has been replaced with a proportionality, because the term prob(data|I) has been omitted.

- Reference: Sivia, D. S., & Skilling, J. (1996). Data analysis: a Bayesian tutorial. *AMC*, *10*, 12.

Comparison of Least-Square Method with Maximum Likelihood Function

For parameter estimation problems with simple linear relation model, Colin and co-workers (2010) pointed out that least-square method is the same as maximum likelihood method when the errors are Gaussian and additive.In the following, a trivial problem will be taken as an example to demonstrate the above theory.

$$y_i = x + \xi_i; \quad (i = 1, 2...N)$$

In the above additive error fitting model, where y_i is the data, x is the parameter to be estimated and ξ_i are the errors.

$$x_{LS} = \sum_{i=1}^{N} y_i/N = \bar{y}$$

With least-square method, the best estimate of y_i is given as the mean as shown above. On the other hand, Maximum likelihood function method is conducted. The likelihood function is given in the following:

$$\begin{aligned} p((y_i)_i = 1, 2, ...N|x) &= p(y_1|x)p(y_2|x)\dots p(y_N|x) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{(-\frac{1}{2\sigma^2}\xi_1^2)} \frac{1}{\sqrt{2\pi}\sigma} e^{(-\frac{1}{2\sigma^2}\xi_2^2)} \cdots \frac{1}{\sqrt{2\pi}\sigma} e^{(-\frac{1}{2\sigma^2}\xi_N^2)} \\ &= \frac{C}{(\sqrt{2\pi}\sigma)^N} e^{-\frac{N}{2\sigma^2}(x-\bar{y})^2} \end{aligned}$$

It is interesting to find out the likelihood function is in Gaussian form and x is given in the following:

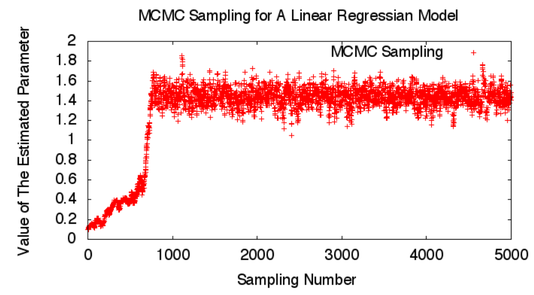
$$x \sim N(\bar{y}, \frac{\sigma^2}{N})$$

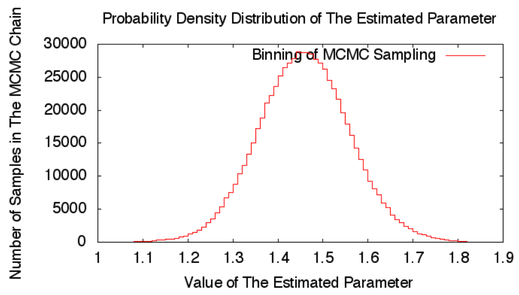
which can give the same results with least-square method. However, when the error is not Gaussian and additive, or the models are not simple linear models, parameter estimation and model selection based on Bayesian inference can give more detailed information, which are more appropriate than least-square method.

- Reference: Colin Fox, Geoff K. Nicholls, Sze M. Tan "An Introduction To Inverse Problems" 2010.

Why Markov Chain Monte Carlo (MCMC)?

In general, MCMC provides a means of sampling ("simulating") from an arbitrary distribution. The purpose of MCMC method is for sampling from probability distributions. A Markov chain is built up to sample the desired distribution. Each point is based on and only based on the last point. Thus, the MCMC is not given a iid (individual independent distribution), which is the price to be paid for using this sampling method.

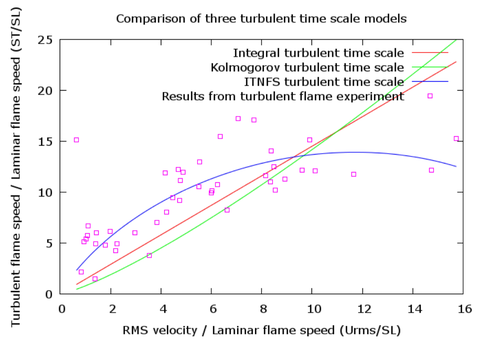




The left side is an example of MCMC sampling. The MCMC method used in CRF is based on Metropolis-Hastings algorithm. After a short term of random walk, the estimated parameter converged into some range. The probability density distribution (PDF) of the estimated value can be obtained from binning of the sampling points.

A forward problem: Parameter estimations of three turbulent time scales based on data of a turbulent flame experiment

On the right side, it is a real forward problem based on Bayesian inference. The data shown by symbols are measured from a turbulent flame experiment. Three turbulent time scale models with additive error are proposed as following:



Integral: $y_i = \alpha x_i + \sigma\xi_i$

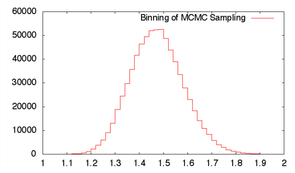
Kolmogorov: $y_i = \alpha(x_i)^{\frac{5}{4}} + \sigma\xi_i$

ITNFS: $y_i = \alpha\sqrt{\Gamma x_i} x_i + \sigma\xi_i$ and $\Gamma(x_i) = A + Be^{-x_i^{1/3}}$

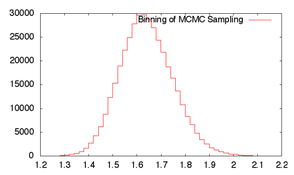
Then, a common way for parameter estimation of forward problems with Bayesian inference is shown in the following flow path:

Build up likelihood function → Sampling with MCMC → Binning to get PDF

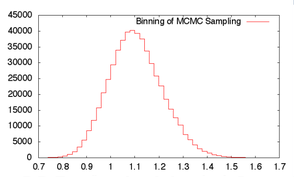
The PDFs of α (and β for model 3) and σ of the three cases can be given through the above process and the best fitting models are demonstrated in the above figure.



PDF of σ in Model 1



PDF of σ in Model 2



PDF of σ in Model 3

From of PDFs of σ, the bias showed by PDF of σ distribution of fitting model 3 is obviously smaller than other two models. However, for a model selection process based on Bayesian inference, the summarized bias is not the only thing need to be considered. There are two estimated parameters in integral model and Kolmogorov model, but there are three estimated parameters in ITNFS model. That illustrates the Ockham factor in fitting ITNFS model will be larger than other two models. Thus, for model selection of above problem, more detailed investigation need to be conducted and Ockham factor need to be considered.

An inverse problem: Data-free inference of uncertain parameters in chemical models

The rate coefficient data for the reaction $H+O_2$ ->OH +O provided in the paper by Masten and Hanson (1990), obtained by fitting kinetic models to experimentally observed OH time histories, are used in performing Bayesian inference of the rate coefficient using a DFI algorithm proposed herein.

Masten and Hanson (1990) report experimental results of shock tube experiments in the form of reaction rates for the chain-branching reaction above at various starting initial temperatures and mixtures. They also report uncertainty associated with the reported values. These results are given in the table.

T_5, K	P_5, atm	ρ ($\times 10^{-5}$), mol cm^{-3}	H_2, mol %	O_2, mol %	k_2 ($\times 10^{12}$), cm^3 mol^{-1} s^{-1}
Incident Shock Waves					
1449	1.457	1.225	2.54	0.492	0.601
1452	1.461	1.226	4.99	0.500	0.577
1518	0.886	0.711	5.04	0.500	0.683
1579	1.000	0.772	4.96	0.492	0.826
1589	0.606	0.465	4.97	0.496	0.850
1635	1.100	0.820	3.18	0.317	0.971
1710	0.461	0.329	2.14	0.212	1.25
1806	0.638	0.431	4.08	0.400	1.50
1813	0.666	0.448	2.55	0.477	1.56
1823	0.708	0.473	4.08	0.400	1.47
1980	0.675	0.416	5.00	0.493	2.15
1986	0.599	0.368	2.14	0.212	2.31
2028	0.479	0.288	4.97	0.496	2.49
2172	0.565	0.317	4.96	0.492	3.00
2206	0.432	0.239	4.97	0.496	3.34
2408	0.672	0.340	3.18	0.317	4.21
2701	0.358	0.162	4.97	0.496	6.22
2734	0.379	0.169	3.18	0.317	6.42
Reflected Shock Waves					
1960	2.525	1.570	1.82	0.181	1.84
2041	2.383	1.423	1.10	0.208	2.36
2105	2.644	1.531	1.945	0.199	2.47
2151	2.442	1.384	0.964	0.200	2.78
2234	2.189	1.194	0.906	0.168	3.23
2298	2.279	1.209	0.906	0.168	3.53
2386	2.345	1.198	1.82	0.181	3.86
2616	2.037	0.949	0.461	0.046	5.39
2681	2.083	0.947	1.014	0.215	5.77
2769	1.991	0.876	0.964	0.200	6.31
2898	1.984	0.834	1.10	0.208	7.11
3370	1.590	0.575	0.709	0.122	10.80

Given an assigned uncertainty of 40% to k_2, k_2 as a random variable will be modeled. More specifically, the pre-exponential factor A_2 will be modeled as a log-normal r.v., while assigning n_2 and E_2 nominal values.

For the DFI problem at hand, the following summary statistics denoted by I:

1. Initial temperatures at which k statistics are available T^i, I = 1, ……, 30
2. Nominal values for k for all initial temperatures
3. Standard deviations for k for all initial temperatures

With the above given information I, the state vector for the outer chain, denoted by ξ and the state vector for the inner chain denoted by λ are demonstrated in the following. Thus, the outer chain will be solving the Bayesian problem

$$p(\xi|I) \propto p(I|\xi)\pi(\xi)$$

and the inner chain solving

$$p(\lambda|I) \propto p(I|\lambda)\pi(\lambda)$$

With the DFI code developed in CRF, the inverse problem can be solved by MCMC sampling.

- Reference: Masten, D. A., Hanson, R. K., & Bowman, C. T. (1990). Shock tube study of the reaction hydrogen atom+ oxygen. fwdarw. hydroxyl+ oxygen atom using hydroxyl laser absorption. *Journal of physical chemistry*, *94*(18), 7119-7128.

Conclusion

1. Bayesian inference is a powerful tool for uncertainty quantification combined with MCMC sampling, both for forward problems and inverse problem.
2. A forward problem is investigated, which shows Bayesian inference can provide a reliable way for parameter estimation and model selection.
3. An inverse problem will be investigated in the summer, which produce a possible way to analyze uncertain parameters in chemical models with missing data.

This work was supported by the US DOE Office of Basic Energy Sciences, Div. of Chem. Sci., Geosci., \& Biosci.
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration