Harnessing Data Sets for More Accurate Kinetic Parameter Estimation: Bayesian Parameter Estimation To Include Errors from Both Experiment and Theory

Aditya “Ashi” Savara¹, Eric A. Walker²

1. Chemical Sciences Division, Oak Ridge National Laboratory,
TN, USA, savaraa@ornl.gov

2. Computational and Data-Enabled Science and Engineering,
State University of New York at Buffalo, NY, USA,
ericwalk@buffalo.edu

CheKiPEUQ

<https://github.com/AdityaSavara/CheKiPEUQ/>

a tool for the future

By A. Savara & E. A. Walker

Talk Outline

- The Big Picture (why do kinetic modeling?)
- Parameter Estimation (what is BPE?)
- CheKiPEUQ (how to do BPE?)
- Bayesian Design of Experiments
- Acknowledgements

The Big Picture

- Perform experiments and calculations for **accurate predictions**, via activation energies, etc.
- We have guesses from previous experiments, density functional theory, linear descriptors, etc. These have some uncertainties. However, even that uncertainty is valuable information! Don't throw it away!*
- What to do??
  - Combine experimental data and computational data to get the most physically accurate values!
  - Comparing to experiments requires doing kinetic modeling. In conjunction with simulations, we will use Bayesian parameter estimation.

parameter estimation by experiment + simulation

Example with Temperature Program

- Adsorb molecules, raise T gradually (e.g., 2 Kelvin per second) monitor the desorption / reaction. Such temperature programmed reaction experiments are common.

- Then, simulate kinetics by functions: **parameters** → **rate**. This is a common way to get activation energies.

But normally, people just vary E_a / A to try to get the best fit.

Conventional Parameter Estimation

- Conventional Parameter Estimation seeks the best fit. However, the best fit is not enough.
- What if Density Functional Theory or chemical knowledge tell us that the activation energy must be close to 40 kJ mol^{-1} .
- What to do if the best fit is 10 kJ mol^{-1} ? What if that is physically unrealistic.?
- Should we penalize unrealistic values during fitting? If so, by how much?

Bayesian Parameter Estimation

Don't try to penalize. Change the question. With Bayesian Parameter Estimation, we change the question from “what is the best fit?” to “what is the most physically probable?”

This has a well-defined solution.

Probability of **D** (observed data) given θ (specific values)

Bayes' Theorem:

$$P(\theta|D) = \frac{P(D|\theta)*P(\theta)}{P(D)} = \frac{P(D|\theta)*P(\theta)}{\text{constant}}$$

Probability of θ (prior, from theory)

Probability of θ (specific values) given **D** (observed data)

D = observed data

θ = vector of parameters (e.g., activation energies, pre-exponential factors, adsorbate interaction terms, etc.)

CPE vs. BPE

- We intentionally use the word ***estimation*** rather than ***optimization*** to emphasize that we are not trying to find the best fit, and instead seek the most physically realistic values.
- **Conventional Parameter Estimation (CPE):**
  - Optimize the “Objective Function” is typically a weighted sum of squared residuals (WSSR).
- **Bayesian Parameter Estimation (BPE):**
  - Find *most probable* set of parameters, based on prior assumptions (i.e., physical beliefs)

Consider TPR with Ethanal/CeO₂(111)

Ethanal desorbing from CeO₂(111), during temperature ramp results in ~ 225 K peak.
Light grey shows estimated uncertainty.

Chen, T. L.; Mullins, D. R. "Adsorption and Reaction of Acetaldehyde over CeO_x(111) Thin Films."
J Phys Chem C **2011**, *115*, 3385

To reproduce the long tail, we first consider a model which uses **two adsorption states**, each with **linearly coverage dependent activation energies**

$$k_i = A_i e^{-\frac{E_{a,i}(\theta_T)}{RT}}$$

$$\theta_T = \theta_1 + \theta_2$$

$$E_{a,i}(\theta)$$

The equations themselves are not the point of this talk.

Some Parameters → Rate

We'll compare CPE & BPE

$$r_T = \sum_i r_i$$

Figure reprinted with permission from: Sebastian Matera, William F. Schneider, Andreas Heyden, and Aditya Savara. *ACS Catalysis* **2019** *9* (8), 6624-6647 Copyright 2019 American Chemical Society

CPE with Objective Function

- Find best fit by changing parameters.
  - Using the error weighted Sum of Squared Residuals (WSSR).
  - We obtain a great fit!

Savara, A. “Simulation and Fitting of Complex Reaction Network TPR: The Key Is The Objective Function” *Surf Sci*, **2016**, 653, 169. DOI: 10.1016/j.susc.2016.07.001

Figure reprinted with permission from: Sebastian Matera, William F. Schneider, Andreas Heyden, and Aditya Savara. *ACS Catalysis* **2019** 9 (8), 6624-6647 Copyright 2019 American Chemical Society

BPE with Bayes Theorem

- Start with physical knowledge of $E_{a,i} = 40 \pm 20$ kJ/mol (DFT) and $A_i = 10^{13 \pm 2}$ s⁻¹ (kinetic theory)
  - Find the maximum of the posterior distribution (**MAP**)

CPE vs. BPE

Why use BPE for worse fit!?

Who says it is worse?!

More realism using BPE

- Blue regions are physically realistic. True values are expected within second blue oval (95%).
  - The BPE values are more realistic than CPE values.
  - What if we go to a more complex E_a dependence?

One Species, Adsorbate Interact.

- Allow nonlinear coverage dependence: $E_a(\theta) = E_a(0) - \gamma(\theta)\theta$
  - Real adsorbate coverages are not simply linear!
- More flexibility enables one species to produce the peak.

$$E_a(0) = 13.6 \text{ kJ/mol}$$
$$A = 6.0 \text{ s}^{-1}$$

$$E_a(0) = 29.75 \text{ kJ/mol}$$
$$A = 14.9 \text{ s}^{-1}$$

Requires only $< 5 \text{ kJ/mol}$ coverage dependence!

But required beyond linear! BPE values more realistic, again.

Closer Look at BPE Results

- The HPD (high probability density region, in red) shows a strong correlation between E_a and A .
- These graphs and the graphs of the previous slide are output by the CheKiPEUQ software.
- The software will create plots of the HPD, correlations, and also simulated data versus the observed data.

Complexity is Good, not Bad

- Once nonlinear adsorbate interactions are allowed, there are often no unique solutions.
  - The real strength of BPE comes for such cases.
  - Why? Because BPE seeks the most realistic solution from all solutions. So it is *ideally suited* to cases that would normally be classified as overfitting.

How to perform BPE?

get CheKiPEUQ

Define the Experimental Data

- For CheKiPEUQ, the data only needs to be imported into a python namespace.
  - We need experimental data & the uncertainties.
- Typically, we just import these from a csv file.

x axis (which can even be observation number)

Experimental Data

```
UserInput.responses['responses_abscissa'] = times
UserInput.responses['responses_observed'] = responses_observed
UserInput.responses['responses_observed_uncertainties'] = observedResponses_uncertainties
```

Enter in the Likelihood

Experimental uncertainties

Build Model

The model is anything that uses parameters to simulation. Arbitrary python code can be used.

- A cantera & csv interface has been built.
- A CATMAP interface is planned.
- A KMC interface is planned.
- Can be even be used for non-kinetics, e.g., molecular spectroscopy identification etc.

	A	B	C	D	E	F
1	reactionID	canteraReactionType	reactionEquation	A	b	E
2	1	surface_reaction	Acetaldehyde + CeCation(S) => Acetaldehyde1-Ce(S)	1	0	2000
3	2	surface_reaction	Acetaldehyde1-Ce(S) => Acetaldehyde + CeCation(S)	1.00E+13	0	67400
4	3	surface_reaction	Acetaldehyde + CeCation(S) => Acetaldehyde2-Ce(S)	1	0	2000
5	4	surface_reaction	Acetaldehyde2-Ce(S) => Acetaldehyde + CeCation(S)	1.00E+13	0	67400
6						

This software is extremely general. It is not limited to chemical kinetics. It can be used for other problems also.

Define the Priors

- For CheKiPEUQ, the standard ways are inside the python runfile, UserInput namespace, or csv files. Use 'arrays' or python pointers.

DFT / Linear Relations etc.

```
UserInput.model['InputParameterPriorValues'] = [41.5, 41.5, 13.0, 13.0, 0.1, 0.1] # Ea1
UserInput.model['InputParametersPriorValuesUncertainties'] = [20, 20, 2, 2, 0.1, 0.1] #
UserInput.model['InputParameterInitialGuess'] = [40, 20, 13, 13, 0.1, 0.1] #This is whe
```

Enter in the Prior

**(optional) guess for where
to start 'search' of parameter space**

**Uncertainties associated
with the prior knowledge**

Seek Complexity

- CheKiPEUQ has been specifically built to handle arbitrary complexity.
- Remember, the real strength of BPE is *not* from trying to solve simple cases. BPE benefits the field more as we intentionally move to the more complex cases.
- Large datasets will give us meaningful priors. The field is well poised for analyzing situations that would be classified as ‘overfitting’ with conventional parameter estimation.

Do the Simulations

- CheKiPEUQ users don't need to learn BPE equations. Just put in the starting values.
- The code will output both BPE and CPE results and even transform data for better searching.
  - Various search types supported: Markov Chain Monte Carlo, conventional optimizers, grid search.
  - Tutorial and guide explain the possible choices.

Get out the Posterior

Beyond Parameter Estimation

Design of Experiments & Model Discrimination

(also in CheKiPEUQ)

Design of Experiments

- Bayesian analysis can also be used for design of experiments. Consider a situation where there is **an equilibrium between 2 different states, separated in energy by ΔG** . We have a prior estimate for ΔG . With a given experimental uncertainty, **what temperature is best to measure at to gain the best estimate of the true ΔG ?**

Model Discrimination

- Bayesian Parameter Estimation can be performed with two models (for example, two possible sites).
- Then, Bayesian Model discrimination to assess *which* model is more probable given data. The simulation ‘closer’ to experiment is **not** always more probable.

Acknowledgements & Future

There is more to learn. Both for you, and for us!

CheKiPEUQ

<https://github.com/AdityaSavara/CheKiPEUQ/>

to be public by mid 2020

A. Savara & E. A. Walker

(will be submitting to *ChemCatChem*)

Work by A. Savara sponsored by the U.S. Department of Energy (DOE), Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences.

U.S. DEPARTMENT OF
ENERGY