Computers and Chemical Engineering 48 (2013) 14–28

Contents lists available at SciVerse ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Estimating reaction model parameter uncertainty with Markov Chain Monte Carlo

Jacob Albrecht*

Bristol-Myers Squibb, One Squibb Drive, New Brunswick, NJ 08901, United States

ARTICLE INFO

Article history:
Received 8 December 2011
Received in revised form 6 July 2012
Accepted 10 July 2012
Available online 31 July 2012

Keywords:
Markov Chain Monte Carlo (MCMC)
Non-parametric statistics
Chemical kinetic modeling
Nonlinear regression

ABSTRACT

Predicting the performance of chemical reactions with a mechanistic model is desired during the development of pharmaceutical and other high value chemical syntheses. Model parameters usually must be regressed to experimental observations. However, experimental error may not follow conventional distributions and the validity of common statistical assumptions used for regression should be examined when fitting mechanistic models.

This paper compares different techniques to estimate parameter confidence for reaction models encountered in pharmaceutical manufacturing, simulated with either normally distributed or experimentally measured noise. Confidence intervals were calculated following standard linear approaches and two Markov Chain Monte Carlo algorithms utilizing a Bayesian approach to parameter estimation: one assuming a normal error distribution, and a new non-parametric likelihood function. While standard frequentist approaches work well for simpler nonlinear models and normal distributions, only MCMC accurately estimates uncertainty when the system is highly nonlinear, and can account for any measurement bias via customized likelihood functions.

1. Introduction

Pharmaceutical and fine chemical process development frequently involves reaction analysis to fully characterize and predict process performance. Using a variety of different analytical tools, scientists collect time-dependent measurements of process parameters (temperature, volume, concentration, etc.) and analyze this data to calculate the underlying fundamental chemical kinetics. The pharmaceutical industry is rapidly adopting the concept of Quality by Design (QbD), wherein the quality of pharmaceutical products is assured by a rigorous understanding of the relationships between manufacturing conditions and final product characteristics (Garcia, Cook, & Nosal, 2008). Regulatory guidance from the Tripartite International Conference on Harmonization (ICH) Q8 (R2) provides definitions of design space and quality attributes for pharmaceutical manufacturing processes. The combination of the model, parameter, and measurement uncertainty defines the size and shape of the process design space (Hallow et al., 2010; Peterson, 2008, 2010). Understanding the rates of desired and undesired reactions is not only good engineering practice, but can define the experimental planning that ultimately determines the acceptable processing ranges for pharmaceutical manufacturing which must be submitted to industry regulators. While it is desired to create a mechanistic model for the system of reactions, it is important to minimize the impact of experimental error on regressed parameters. Often in pharmaceutical process development, time and reagents are limited. Therefore, a finite amount of data is collected during process development and the resulting model parameters have an inherent uncertainty that must be considered when proposing a model-based process design.

In the case of chemical concentrations, kinetic data can be collected from tools such as high performance liquid chromatography (HPLC) or in-line Fourier transform infrared (FTIR) spectroscopy by first preparing concentration standards and then correlating the analytical output to known concentrations. For HPLC this calibration curve conventionally uses a linear correlation between chromatogram peak area and concentration. Industrial guidance, for example from the United States Pharacopeal Convention (USP) Chapter 1225, provides a definition of linearity for an analysis method, ideally for results proportional to concentration, but does allow for "well-defined transformations" of data for analytical techniques if warranted. For FITR, a chemometric model is built using partial least squares (PLS) regression of infrared spectra to concentration (Kramer, 1998). Once these machine calibrations are created, unknown samples can be analyzed. The measurement error of these samples may be assumed to be due to the inherent error of the calibration as long as the sample preparation and machine configuration are consistent between samples. Unfortunately, once these calibrations are created and deemed acceptable, additional information contained in the goodness of fit and model

* Corresponding author. Tel.: +1 732 227 6330.
E-mail address: jacob.albrecht@bms.com

0098-1354/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compchemeng.2012.07.011

Nomenclature

Variables and functions

a	metropolis acceptance probability
$[A]$	concentration of chemical A
$[B]$	concentration of chemical B
$[C]$	concentration of chemical C
$[D]$	concentration of chemical D
$E()$	expectation value
$\mathbf{F}$	linearized design matrix
$F^{\alpha}_{p,n-p}$	F-statistic
$f()$	model for concentration as a function of time
$g()$	indicator function
$L()$	likelihood function
$lookup()$	lookup table query
$N()$	normal distribution
n	number of measurements
N_s	total number of Monte Carlo iterations
$N_{bins,j}$	number of histogram bins for parameter j
N_{chain}	number of simultaneous Markov chains (2)
p	number of parameters
$p()$	probability density function
R	ideal gas constant
$S()$	sum of squares error
T	temperature
T_{ref}	reference temperature (298 K)
t	time
$t^{\alpha/2}_{n-p}$	Student's t-distribution significance level
$U()$	uniform distribution
u	random sample on $U(0, 1)$
w	weighting factor
$\mathbf{Y}$	vector of chemical concentration measured over time

Subscripts and superscripts

0	concentration at time zero
b	histogram bin index ($b = 1, 2, \ldots, N_{bins}$)
i	timepoint measurement index ($i = 1, 2, \ldots, n$)
j	parameter index ($j = 1,2, \ldots, p$)
LS	least squares optimum
m	model index (1, 2, 3, 4)
k	Monte Carlo iteration ($k = 1 \ldots N_s$)
$optimum$	optimum value that maximizes the model likelihood function
$prop$	proposed value
$true$	exact observation without error

Greek variables

α	confidence level for t or F distribution
ε	random error
Δbin	width of histogram bin
σ	standard deviation
$\boldsymbol{\theta}$	vector of parameters in model
$\boldsymbol{\Theta}$	matrix of proposed parameter values accepted by MCMC

residuals is typically ignored. If the majority of experimental error can be determined by preparing and analyzing replicate samples, there is an opportunity to use this measurable error in kinetic parameter regression.

Mechanistic or empirical reaction models can be built by regressing a kinetic data set to a proposed system of equations using linear or nonlinear regression techniques (Dowdy, Weardon, & Chilko, 2005; Seber & Wild, 2005) typically involving least squares regression. In the development of least squares regression techniques and their associated confidence intervals, it has been assumed that errors were independent and identically distributed (Dowdy et al., 2005). Because measurement error can be determined using replicate samples of reference standards, and adequate models can often be selected when the reaction mechanism is known, much of the overall uncertainty lies in the quality of parameter estimation. Frequently, the error of a measurement technique is described not in absolute concentration, but in terms of percent, an example of heteroscedastic error. For mean-centered heteroscedastic error, weighted least squares is commonly used when the error distribution is known. Once the best fit parameters are found using ordinary or weighted least squares, basic parameter confidence calculations rely on a Taylor series linearization of the model at this best fit condition. Despite abundant warnings on using linearization methods to estimate confidence intervals in the literature, their prominence in many statistical packages encourages their use for potentially inappropriate models. Increasingly though, these assumptions for nonlinear chemical kinetic models are being challenged, but only a few authors have empirically determined their accuracy on a case-study basis (Donaldson & Schnabel, 1987; Sin, Meyer, & Gernaey, 2010).

Estimating parameter uncertainty for nonlinear models has been studied extensively by other authors. Approaches can generally be collected into two groups. The classical approach developed by statisticians such as Fisher, Pearson, Neyman (Neyman, 1937), and others to determine confidence that an observation can be repeated from a given model will be referred to here as the frequentist approach. This general approach leads to several well-established methods for measuring parameter confidence, including now-ubiquitous linear least squares approaches. For highly nonlinear models, other techniques such as model transformations, scaling, likelihood maximization, likelihood ratio, and lack of fit methods can be used to more accurately calculate confidence regions compared to linear or quadratic approximations (Bates & Watts, 1981; Donaldson & Schnabel, 1987; Seber & Wild, 2005), but these techniques require a closed form model to manipulate. Rooney and Biegler (2001) applied the likelihood ratio test to solve chemical engineering design problems. Confidence intervals for nonlinear parameter estimation have been used for activation energy estimation from experimental data, using corrections to Student's t-distribution (Cai, Han, Chen, & Chen, 2011; Vyazovkin & Sbirrazzuoli, 1997). Monte Carlo approaches have also been used to determine regression confidence in chemical kinetic parameters by repeatedly randomizing and regressing data (Alper & Gelb, 1990). A model for cellulose hydrolysis was analyzed by Monte Carlo to determine the uncertainty of model predictions (Sin et al., 2010). Such rigorous analyses are required to determine if a model is fit-for-purpose when parameters cannot be uniquely estimated from experimental data.

As an alternative to frequentist methods, the Bayesian approach aims to determine the credibility of a model given an observation. This approach has been applied to chemical kinetic models (Box & Tiao, 1973) and multiple-response data sets (Stewart, Caracotsios, & Sorensen, 1992; Stewart, Shon, & Box, 1998). To evaluate the credibility of model parameters, Markov Chain Monte Carlo (MCMC) has emerged as a powerful solution method for Bayesian inference by calculating parameter likelihood on a random walk through parameter space. MCMC using unbiased heteroscedastic error expressions for chemical reaction data has been applied to model selection (Blau et al., 2008; Hsu et al., 2009). Coleman and Block (2006) exploited informative priors to estimate fermentation model parameters using MCMC. Klinke (2009) demonstrated the use of adaptive MCMC (Haario, Saksman, & Tamminen, 2001) to evaluate complex cellular signaling networks, simultaneously regressing 34 parameters given experimental data. These methods using MCMC with