Inferring Chemical Reaction Rates from a Sequence of Infrared Spectra

by

Peter Starszyk

A thesis
presented to the University Of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2016

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Certain excerpts in this thesis containing chemistry insights are attributed to direct discussions with the collaborator, Dr. Hind A. Al-Abadleh; such excerpts do not have a reference cited but are stated to have been verified. All experimental data presented in this thesis has been provided by Dr. Hind A. Al-Abadleh and her research team at the Faculty of Science, Wilfrid Laurier University, Waterloo, Ontario, Canada.

Abstract

Many chemical compounds used by the energy and agricultural industries introduce large amounts of arsenic into the environment. As this poses serious health and environmental risks, designing safe and effective decontaminating agents remains an active research area. To do this, it is crucial to understand the chemical kinetics between arsenic and certain geochemicals at the molecular level; of particular interest are the reaction rate constants which describe the behaviour and properties of arsenic in relation to different chemicals. These rates can be inferred from a time series of individual concentration measures of all constituent chemicals in a mixture. However, current laboratory technology cannot produce such measures but instead produces time series of infrared spectra, from which individual chemical concentrations must be deconvoluted. Existing techniques to analyze such data focus on minimizing modeling assumptions and point estimation. In this thesis, we propose a fully specified parametric statistical model directly relating the rate constants to the spectra. This model drastically reduces the number of free parameters, offers statistically principled uncertainty estimates for parameters of interest and provides the added flexibility of incorporating important prior information, which current methodologies do not seem to account for. We further apply the model to experimental data in order to compare two plausible models of arsenic neutralization.

Acknowledgements

First and foremost, I must express the utmost gratitude to my supervisor, Professor Martin Lysy, whose generosity in offering time and brilliant ideas has been nothing short of an inspiration. Both his classroom teaching approach and collaborative attitude towards research have tested my problem solving creativity at ambitious levels. From him, I have indirectly learned that although scientific problems may not have ultimate solutions, they instead have countless steps to be taken forward.

A special thank you goes to Dr. Al-Abadleh. Her insights and enthusiasm for chemistry have been, and continue to be, of great value to this research. I would also like to thank the committee members, Professors Paul Marriott and Greg Rice for offering their time and input.

After taking so many Statistics courses, I wish to thank all of the phenomenal professors that I have had the privilege of learning from and interacting with over the years. I would particularly like to express my appreciation to Professor Pengfei Li whose extraordinary teaching and one-to-one statistical discussions are what initially sparked my genuine interest in the field. Had I not taken his course in the Spring of 2013, this thesis would certainly have been written by a different student in a different time.

I am grateful to Mary Lou Dufton and Leanne Bird, who not only work tirelessly to make life easier for the entire department, but are amazing people overall. I would also like to thank my friends in the department: Mirabelle, Marco, Garcia for making the graduate experience all the more enjoyable. I would further like to thank my fellow classmates, namely Gary Song for always offering new mathematical perspectives and Lin Qin, my “Bayes” friend.

I am fortunate to have great parents, grandmother and sister who offer their unconditional support. Last but not least, I thank Jennifer for her patience and care, you have always been there for me.

Niniejszą pracę dedykuję moim Rodzicom.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and Related Work

Arsenic is an element naturally found in minerals and rocks all around the world. In addition to its natural formations, strongly contributing to its presence are the biogeochemical processes provoked by industry such as biological pretreatment of solid waste, coal combustion and the use of herbicides and pesticides [1, 2]. As the increased presence of arsenic poses serious health and environmental risks [39] and thus challenges for growing industry, designing safe and effective decontaminating agents remains an active research area.

In order to design such decontaminating agents it is important to understand the chemical kinetics between arsenic and different geochemicals. In this thesis, we focus on one particular study conducted by a chemistry research group led by Dr. Hind A. Al-Abadleh at Wilfrid Laurier University. This study seeks to understand the chemical reactions that occur between Dimethylarsinic acid (DMA; otherwise denoted as species S_1) and iron oxide. The group has published a number of experimental studies using infrared spectroscopy complemented with computational chemistry results [33, 34, 35, 36, 37]; this body of published work revealed that the surface chemistry of DMA proceeds by forming three types of surface species with iron oxide (species S_2, S_3, S_4; details in 2.2). A system of Ordinary Differential Equations (ODEs) describes the concentrations $X_t = (X_{1t}, X_{2t}, X_{3t}, X_{4t})$ of each species at time t as a function of unknown reaction rate constants, κ, and initial concentrations, X_0 (details in 2.1). There is a considerable body of statistical literature on estimating ODE parameters from fully or partially observed components of X_t at discrete time points [31, 32]. However, our data consists of a sequence of infrared spectra; the study of the interaction between molecules and the infrared region of the electromagnetic spectrum.

In particular, our data represents the infrared photon absorption measures of a chemical mixture (details in 3.2) over experimental wavenumbers $i \in \{1, 2, \ldots, n\}$ across experimental time points $t \in \{1, 2, \ldots, T\}$. By the Beer-Lambert Law [3], the absorption $\mathcal{A}_t(\omega_i)$ of a mixture at time t for a particular infrared wavenumber ω_i is $\mathcal{A}_t(\omega_i) = \sum_{j=2}^{4} \mathcal{A}_{jt}(\omega_i)$ where $\mathcal{A}_{jt}(\omega_i)$ is the wavenumber-specific absorption of one mole of each individual species $j = 2, 3, 4$. The challenge is to disentangle the ODE parameters (κ, X_0) from the infrared spectra; in particular our parameters of interest are the reaction rate constants κ.

To currently do this, the Multivariate Curve Resolution (MCR) [6, 7] method is widely used in chemometrics. MCR aims to reconstruct the absorption data reasonably well by an additive bilinear function; a linear combination of spectral components and concentration profiles of constituent species in the mixture [6, 8]. In particular, given an experimental absorption data matrix, $\tilde{\mathcal{A}} \in \mathbb{R}^{T \times n}$, with rows corresponding to experimental time points $t \in \{1, 2, \ldots, T\}$ and columns corresponding to experimental wavenumbers $i \in \{1, 2, \ldots, n\}$, MCR decomposes the data matrix as the product of two unknown matrices plus error [9]

$$\tilde{\mathcal{A}} = \mathbf{C}\mathbf{S}^\top + \mathbf{E}$$

where $\mathbf{C} \in \mathbb{R}^{T \times k}$ is a matrix of concentration profiles of the k constituent species in the mixture, $\mathbf{S}^\top \in \mathbb{R}^{k \times n}$ is a matrix of pure spectral profiles, both of which are estimated from the data, and $\mathbf{E} \in \mathbb{R}^{T \times n}$ is the residual error matrix. The integer k is either known a priori or estimated using a suitable data reduction technique such as Principal Component Analysis (PCA) or Independent Component Analysis (ICA) [6]. Further, the component matrices are often estimated by an Alternating Least Squares (ALS) algorithm [10] subject to chemical plausibility enforcing constraints such as non-negativity of concentrations.

$$\underset{C}{argmin}\left\|\hat{\mathcal{A}}_{PCA} - \mathbf{C}\hat{\mathbf{S}}^{\top}\right\| \leftrightarrows \underset{S^{\top}}{argmin}\left\|\hat{\mathcal{A}}_{PCA} - \hat{\mathbf{C}}\mathbf{S}^{\top}\right\|$$

By and large, the literature and software on this topic has focused extensively on minimizing assumptions and constraints on the theoretical spectra $\mathcal{A}$ and error distribution, and perhaps consequently, on point estimation of reaction rate constants.

Although this method is widely used, it has some disadvantages. The first disadvantage is the rotational ambiguity problem which leads to non uniqueness of solutions for $\mathbf{C}, \mathbf{S}^{\top}$ in the optimization problem [11, 12]. The second disadvantage is that due to minimal assumptions, $\mathbf{C}, \mathbf{S}^{\top}$ are often model free and as such the method optimizes over a parameter space which scales to the size of the dataset, thus optimizing over a $k \times (T + n)$ - dimensional parameter space may be problematic for large T, n, k. Moreover, the non-parametric nature of MCR limits it to offering only point estimates of individual contributions but does not offer parametric interpretations of the Infrared Spectroscopy process nor any statistical information for related parameters of interest.

1.2 Contribution

Due to instrumental and methodological limitations, the challenge is to determine 1) the set of reaction channels governing the chemical system of interest and 2) the corresponding reaction rate constants. The primary contribution in this thesis is the proposal of a model that relates the chemical reaction rate constants directly to the Infrared Spectroscopy process. In particular, we embed the basic bilinear equation underpinning MCR into a fully specified, parametric statistical model of the spectra of each species, thereby reducing the number of free parameters in the model. Bayesian Inference is adopted for parameter estimations which provide statistically principled uncertainty estimates for the parameters of interest and allows for the incorporation of important prior information such as relative magnitudes between rate constants and final concentrations. As a secondary contribution, we apply our proposed model to the experimental data to determine a set of reaction channels that are likely to be governing the chemical system of interest.

1.3 Outline

We first describe the chemical system (sometimes referred to as mixture) being studied and define the candidate reaction systems (with corresponding ODE formulations) which are strongly believed to govern the mixture. We then explain the challenges of inferring rate constants from currently available concentration measures and discuss alternative data used for inference which comprises of an experimental set and a theoretical set. We then discuss the proposed parametric statistical model in detailed layers, followed by a discussion of the Bayesian inference approach used for parameter estimations of the proposed model. After illustrating a simulation study, we apply the model to real data. In particular, the model is applied under two separate sets of reaction assumptions that are believed to govern the system. After both models have been estimated from the experimental data, we compare both model fits and discuss which reaction model is more plausible for the mixture, given the data.

Chapter 2

Chemical Framework

2.1 Review of Chemical ODE Systems

Consider some chemical system composed of d chemical species $(S_1, S_2, \ldots, S_d)$ governed by some set of m reaction channels between the constituent species, each occurring at a particular reaction rate constant $(\kappa_1, \kappa_2, \ldots, \kappa_m)$. Assuming that the mixture is well stirred (all species are uniformly distributed within the mixture), a system of d ordinary differential equations (ODEs) parameterized by m reaction rates can be obtained from these reaction channels to describe the rate of change of the concentration of each species in the chemical system over time $(X_{1t}, X_{2t}, \ldots, X_{dt})$. Moreover, given the initial concentrations of all d species at $t = 0$, the time evolution of all d concentrations (hence chemical presence of each species) is completely determined [13].

As an example, the arsenic system studied (discussed in more detail in Section 2.2) is composed of four species, (S_1, S_2, S_3, S_4). Based on computational chemistry studies, it is strongly believed that the system is governed by one of two candidate sets of reactions. One of them consists of four reaction channels with corresponding reaction rate constants:

$$R_1 : S_1 \xrightarrow{\kappa_{12}} S_2$$

$$R_2 : S_1 \xrightarrow{\kappa_{13}} S_3$$

$$R_3 : S_2 \xrightarrow{\kappa_{23}} S_3$$

$$R_4 : S_3 \xrightarrow{\kappa_{34}} S_4$$

Each reaction corresponds to a molecular mechanism by which the reactant species (LHS) undergo a reaction to produce the resulting intermediates or products (RHS) [14]. As an example, R_1 corresponds to a depletion of 1 S_1 molecule and addition of 1 S_2 molecule at a rate proportional (by κ_{12}) to the concentration of S_1.

To construct the ODE system, we sum the contribution of each species at each reaction separately for any given time t.

i **$\mathbf{S_1}$**. By R_1, S_1 has a net loss of 1 unit at rate proportional (by κ_{12}) to the concentration (at time t) of the species that it requires to deplete S_1; $-\kappa_{12}X_{1t}$. Similarly by R_2, S_1 has a net loss of 1 unit at rate proportional (by κ_{13}) to the concentration (at time t) of the species that it requires to deplete S_1; $-\kappa_{13}X_{1t}$. Thus, the rate of change of total S_1 concentration at any given time is described as

$$\frac{d}{dt}X_{1t} = -(\kappa_{12} + \kappa_{13})X_{1t}$$

ii **$\mathbf{S_2}$**. By R_1, S_2 has a net gain of 1 unit at a rate proportional (by κ_{12}) to the concentration of the species that it requires to create S_2; $\kappa_{12}X_{1t}$. By R_3, S_2 has a net loss of 1 unit at a rate proportional to (by κ_{23}) the concentration of the species it requires to deplete S_2; $-\kappa_{23}X_{2t}$. Thus, the rate of change of S_2 concentration at any given time is described as

$$\frac{d}{dt}X_{2t} = \kappa_{12}X_{1t} - \kappa_{23}X_{2t}$$

iii **$\mathbf{S_3}$**. By R_2, S_3 has a net gain of 1 unit at a rate proportional (by κ_{13}) to the concentration of the species that it requires to create S_3; $\kappa_{13}X_{1t}$. By R_3, S_3 has a net gain of 1 unit at a rate proportional (by κ_{23}) to the concentration of the species that it requires to create S_3; $\kappa_{23} X_{2t}$. By R_4, S_3 has a net loss of 1 unit at a rate proportional (by κ_{34}) to the concentration of the species that it requires to deplete S_3; $-\kappa_{34} X_{3t}$. Thus, the rate of change of S_3 concentration at any given time is described as

$$\frac{d}{dt} X_{3t} = \kappa_{13} X_{1t} + \kappa_{23} X_{2t} - \kappa_{34} X_{3t}$$

iv **$\mathbf{S_4}$**. By R_4, S_4 has a net gain of 1 unit at a rate proportional (by κ_{34}) to the concentration of the species that it requires to create it; $\kappa_{34} X_{3t}$. Thus, the rate of change of S_4 concentration at any given time is described as

$$\frac{d}{dt} X_{4t} = \kappa_{34} X_{3t}$$

Combining the rates of change in the concentrations of each species results in the following ODE system:

$$\begin{aligned}
\frac{d}{dt} X_{1t} &= -(\kappa_{12} + \kappa_{13}) X_{1t} \\
\frac{d}{dt} X_{2t} &= \kappa_{12} X_{1t} - \kappa_{23} X_{2t} \\
\frac{d}{dt} X_{3t} &= \kappa_{13} X_{1t} + \kappa_{23} X_{2t} - \kappa_{34} X_{3t} \\
\frac{d}{dt} X_{4t} &= \kappa_{34} X_{3t}
\end{aligned}$$

The solution (concentration quantities) to chemical ODE systems at any given time $t > 0$, $\{X_{1t}, X_{2t}, \ldots, X_{dt}\}_{t=1}^{T}$, requires integrating the system. In the general case when the solution is analytically unattainable, it can instead be solved numerically; the Runge-Kutta methods are widely used for numerical integration of chemical ODEs [15].

2.2 Description of Chemical Experiment

An experiment is controlled under which the behaviour between DMA (species S_1) and a fixed iron-oxide surface is studied. At $t = 0$, the system is initiated with some quantity of S_1 and three resulting chemical bonds have been verified to occur between the arsenic and the surface. At $t = 0$, the instance arsenic has been initialized, no arsenic molecule has yet come into contact with the surface thus *no bonds* have yet been formed. As time progresses, higher order bonds are formed: *weak bonds, single bonds, double bonds.* The four aforementioned bond states are referred to as species S_1, S_2, S_3, S_4 respectively and we have verified that the system dynamics are strongly believed to be governed by one of the following two candidate reaction systems. Each system describes a set of reaction channels with corresponding reaction rate constant vectors κ:

$$\begin{aligned} &System : 4R \\ &\quad R_1 : S_1 \xrightarrow{\kappa_{12}} S_2 \\ &\quad R_2 : S_1 \xrightarrow{\kappa_{13}} S_3 \\ &\quad R_3 : S_2 \xrightarrow{\kappa_{23}} S_3 \\ &\quad R_4 : S_3 \xrightarrow{\kappa_{34}} S_4 \end{aligned} \tag{2.1}$$

$$\begin{aligned} &System : 3R \\ &\quad R_1 : S_1 \xrightarrow{\kappa_{12}} S_2 \\ &\quad R_2 : S_2 \xrightarrow{\kappa_{23}} S_3 \\ &\quad R_3 : S_3 \xrightarrow{\kappa_{34}} S_4 \end{aligned} \tag{2.2}$$

As the 4 reaction system is more general than the 3 reaction system, we will refer to the former system throughout the remainder of the paper unless stated otherwise.

Throughout the experiment, the interest lies in studying the rate constants at which the molecule reacts with the fixed chemical surface. The nature of the reactions are described as follows:

$\mathbf{R_1}$: S_1 (DMA molecule which has not bonded yet) forms a weak bond with the chemical surface. A weak bond results when the attraction between the molecule and the surface is strong enough to hold but no chemical connection has been made. The formation of the weak bond results in forming outersphere surface species at rate κ_{12}, which is referred to as S_2.

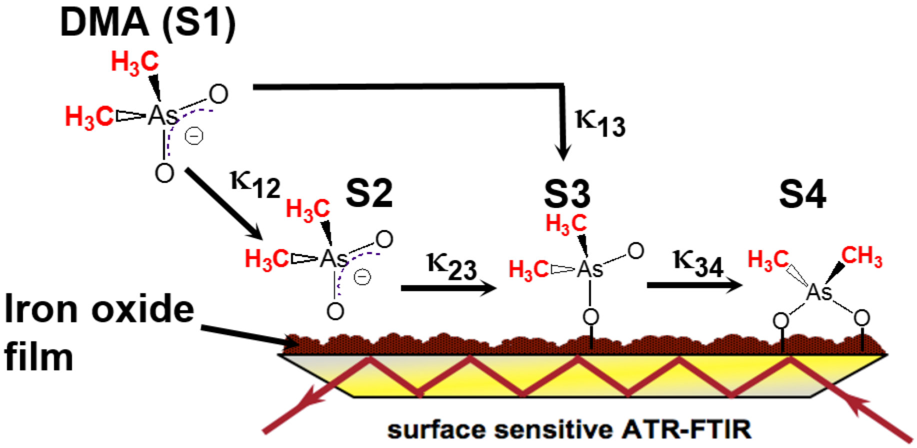
$\mathbf{R_2}$: S_1 directly forms a single chemical bond with the chemical surface. A single bond results when the attraction between the molecule and the surface is strong enough to form a connection between one of the molecule $As - O$ bonds and the chemical surface. The formation of the single bond results in forming a monodentate surface species at rate κ_{13}, which is referred to as S3.

$\mathbf{R_3}$: The outersphere surface species already in a weak bond state transitions into a monodentate surface species with a single bond at rate κ_{23}.

$\mathbf{R_4}$: The monodentate surface species already in a single bond state transitions into a bidentate surface species with double bonds with the surface through the second $As - O$ group of DMA. The formation of S_4 proceeds at rate κ_{34}.

The strength of each bond is significantly greater than any preceding bond ($S_1 \prec S_2 \prec S_3 \prec S_4$). As such it is assumed that no backward reactions occur ($S_i \nrightarrow S_{i'}; \forall i' < i$). In particular, the above reaction channels are chemically referred to as first-order forward reactions; each reaction results in a loss of one $S_{i'}$ bond and a gain of one S_i bond for $i > i'$ [16]. Figure 2.1 illustrates the four reaction process.

Figure 2.1: Sequence of 4 Reaction System



We assume that the mixture is well stirred. As discussed in Section 1.3, inspecting the above reaction channels enables one to describe the rate of change in concentration per unit time of each species in the system by the following set of first order linear Ordinary Differential Equations:

$$\begin{aligned}
&System : 4R\\
\frac{d}{dt}X_{1t} =& -(\kappa_{12}+\kappa_{13})X_{1t}\\
\frac{d}{dt}X_{2t} =& \kappa_{12}X_{1t}-\kappa_{23}X_{2t}\\
\frac{d}{dt}X_{3t} =& \kappa_{13}X_{1t}+\kappa_{23}X_{2t}-\kappa_{34}X_{3t}\\
\frac{d}{dt}X_{4t} =& \kappa_{34}X_{3t}
\end{aligned} \tag{2.3}$$

$$\begin{aligned}
&System : 3R\\
\frac{d}{dt}X_{1t} =& -\kappa_{12}X_{1t}\\
\frac{d}{dt}X_{2t} =& \kappa_{12}X_{1t}-\kappa_{23}X_{2t}\\
\frac{d}{dt}X_{3t} =& \kappa_{23}X_{2t}-\kappa_{34}X_{3t}\\
\frac{d}{dt}X_{4t} =& \kappa_{34}X_{3t}
\end{aligned} \tag{2.4}$$

Given reaction rates $\kappa = (\kappa_{12}, \kappa_{13}, \kappa_{23}, \kappa_{34})$ and initial concentrations $X_0 = (X_{1,0}, X_{2,0}, X_{3,0}, X_{4,0})$, the concentration solution of the ODE system, $X_t = \{X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}\}$, is completely determined for any $t > 0$. Alternatively, X_t is interpreted to be the true model based concentration levels of $\{S_1, S_2, S_3, S_4\}$ at time t, given (κ, X_0). Under the framework discussed at the beginning of the section, $X_{1,0} > 0$ is the true initial concentration of S_1 and $X_{2,0} = X_{3,0} = X_{4,0} = 0$ as no bonds have formed yet.

Chapter 3

Data

3.1 Limitations of Aggregate Concentration Data

As chemical reaction rate constants directly imply relative changes in chemical quantities with respect to time, they can be directly estimated from data which measures the time evolution chemical concentrations of all individual species in the system; $\{Y_{1t}, Y_{2t}, Y_{3t}, Y_{4t}\}_{t=1}^{T}$ (Y_t are the noisy concentrations observed at time t from experimentation). However, limitations in laboratory technology enable only measurements of aggregate concentrations. Since the Beer-Lambert law states that absorption is proportional to concentration, the aggregate concentration is approximated from experimental spectral absorption data (details discussed in 3.2) and only aggregated chemical *concentration data* is attainable $\{Y_{At} = Y_{2t} + Y_{3t} + Y_{4t}\}_{t=1}^{T}$. Figure 3.1 illustrates the observed aggregated concentrations obtained from experimentation (with error bars) and Figure 3.2 illustrates the simulation of plausible concentration levels of each species that the aggregate curve may be composed of.

Figure 3.1: Experimental Aggregate Concentrations

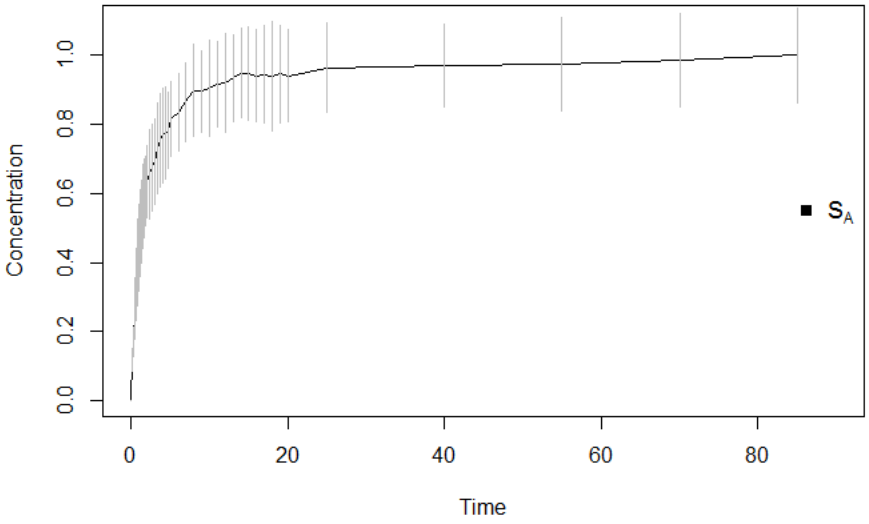
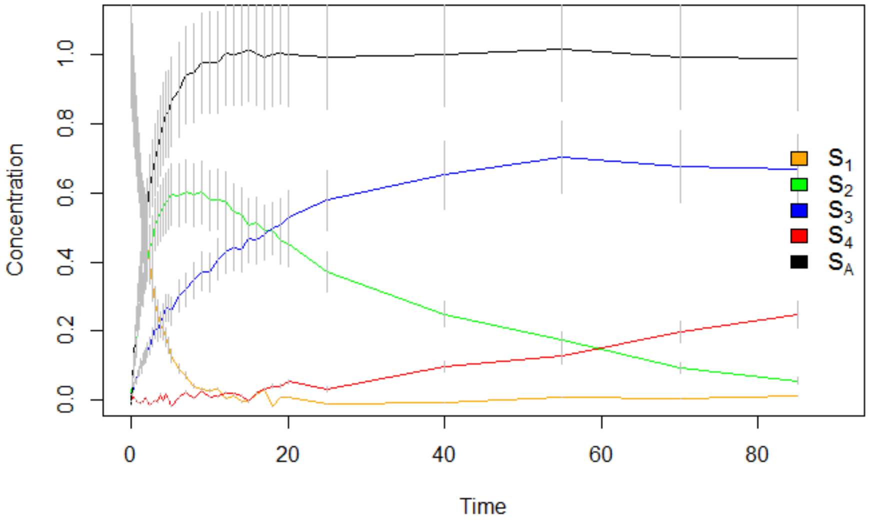


Figure 3.2: Simulated Individual Concentrations



It turns out that the aggregate data, $\{Y_{At}\}_{t=1}^{T}$, is insufficient to infer $\kappa = (\kappa_{12}, \kappa_{13}, \kappa_{23}, \kappa_{34})$ [11]. To understand this, we define $X_{At} = X_{2t} + X_{3t} + X_{4t}$ which simplifies the ODE system (3) to (See Appendix A for derivation):

$$\begin{aligned} \frac{d}{dt}X_{1t} =& -(\kappa_{12} + \kappa_{13})X_{1t} \\ \frac{d}{dt}X_{At} =& (\kappa_{12} + \kappa_{13})X_{1t} \end{aligned} \tag{3.1}$$

in turn reducing the corresponding reaction channels (1) to

$$S_1 \xrightarrow{\kappa_{12}+\kappa_{13}} S_A \tag{3.2}$$

If we choose some $\kappa' = (\kappa'_{12}, \kappa'_{13}, \kappa'_{23}, \kappa'_{34})$ such that $\kappa'_{12} + \kappa'_{13} = \eta$ for some fixed aggregate reaction rate constant $\eta \in \mathbb{R}^+$, we can see, by inspection of (5), that the values of $\kappa'_{23}, \kappa'_{34}$ are completely arbitrary and do not effect the evolution of X_{At}. We can further choose $\kappa'' = (\kappa''_{12}, \kappa''_{13}, \kappa''_{23}, \kappa''_{34})$ such that $\kappa''_{12} + \kappa''_{13} = \eta$ and again the evolution of X_{At} remains unchanged. This suggests that there exists an uncountably infinite set of feasible values for κ that can govern the evolution of the system for some fixed aggregate reaction rate constant $\eta \in \mathbb{R}^+$, and thus the individual reaction rate constants, κ, cannot be uniquely determined given only the aggregated concentration measures. As such, we require additional data for inference.

3.2 Infrared Spectroscopy

As direct aggregate concentration measures from experimental spectral data is insufficient to infer the parameters of interest, chemists turn instead to theoretical and empirical evidence to understand the kinetics. The theoretical component refers to computational chemistry where surface reactions are simulated using model cluster chemicals that mimic the real ones used in the lab. The empirical component refers to experimentation on a chemical system and collecting Infrared (IR) Spectroscopy data. IR Spectroscopy is the study of interactions between molecules and the Infrared region of the electromagnetic spectrum [18].

In particular, the interactions are measured by analyzing the patterns in which a molecule vibrates in response to the IR light [4, 5], which chemically implies the IR photon absorption by that molecule.

3.2.1 Absorption Process

The intensity at which a chemical bond absorbs IR light partly depends on the vibrational frequency of that bond. Two primary modes of vibrations at which IR absorption occurs and are commonly analyzed for kinetic data are stretching and bending of the bond. Recalling that S_2, S_3, S_4 have weak, single, and double bonds, respectively, between the DMA molecule and the iron oxide surface, each species has its own characteristic $As-O$ vibrational pattern in the surface DMA. The aggregate absorptions of $S_A = S_2 + S_3 + S_4$ is measured by the IR spectrometer, where a beam of IR light of a range of wavenumbers $(400-4000cm^{-1})$ is applied to the sample and the IR spectrometer measures the intensities at which S_A absorbs photons across that entire range at predetermined time points after initiation.

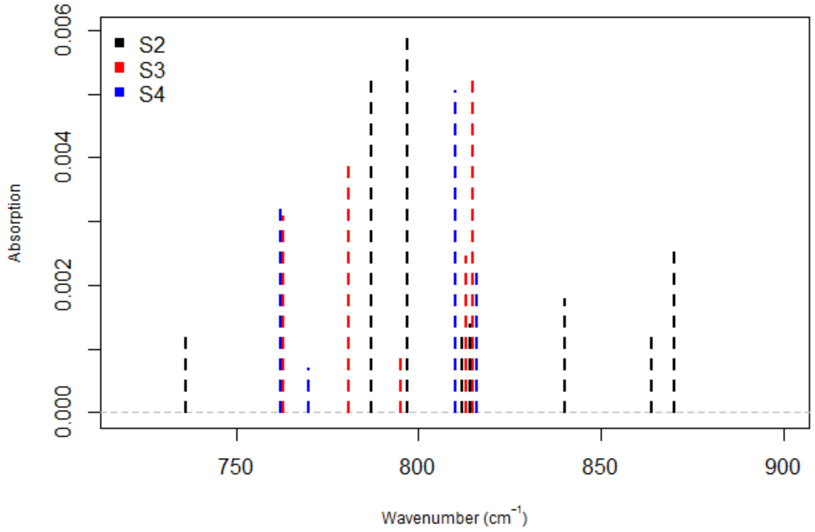
In a typical experiment using ATR-FTIR (Attenuated Total Reflectance - Fourier Transform Infrared Spectroscopy), the flow cell contains the iron oxide film and H_2O as the background solution. The IR intensity of this system, in the absence of DMA (S_1), is recorded and referred to as the "reference spectrum, I_R". Then, a solution of known amount of (S_1) is introduced to the flow cell and the IR spectrometer collects spectra as a function of flow time of S_1. The concentration of S_1 is chosen such that the intensity of IR absorptions is very low and undetectable compared to S_2, S_3, S_4. Hence, throughout the experiment while S_1 is flowing across the iron oxide film, IR intensities of only $S_A = S_2 + S_3 + S_4$ are recorded at given times and referred to as the "sample spectrum, I_S". The final absorption quantity for all wavenumbers ω_i, $i \in \{1, 2, \ldots, n\}$ at times $t \in \{1, 2, \ldots, T\}$, $\tilde{\mathcal{A}}_{it}$, is calculated as, $\tilde{\mathcal{A}}_{it} = \log\left(\frac{I_R}{I_S}\right)$.

3.2.2 Theoretical and Absorbance Spectra

Computational chemistry methods are commonly used in geochemical research and are based on theoretical understanding of the nature of chemical bonds. In general, every chemical species has unique properties which distinguish it from other species. One distinguishing factor pertains to the absorption patterns a species has with respect to particular IR frequencies (expressed in wavenumbers). The unique molecular structure of chemical species causes it to absorb significant amounts of IR photons of particular wavenumbers and not so much of others. The wavenumbers at which the species absorbs significantly are referred to as the "theoretical wavenumbers" and the amounts absorbed at those wavenumbers are referred to as "theoretical intensities". MCR literature does not incorporate this theoretical information but is reflected in our model (details in section 4).

For the system under study, computational chemistry simulations provide a total of 17 theoretical wavenumber/intensity [35] pairs which correspond to the system: 8 belong to S_2, 5 belong to S_3 and 4 belong to S_4 (we note these as $|S_2| = 8, |S_3| = 5, |S_4| = 4$). Figure 3.3 illustrates the locations of the 17 frequencies with heights indicating their relative absorption intensities [35].

Figure 3.3: Theoretical Data



Further, eight experiments are conducted which we assume are independent of one another. For each experiment $l \in \{1, 2, \ldots, 8\}$, a range of 86 equally spaced IR wavenumbers ($671.12cm^{-1} - 998.96cm^{-1}$) are applied to the combined S_A mixture over a very short time interval, each at nine different time points $t = \{1, 2, 3, 4, 5, 10, 15, 20, 85\}$ (mins); we assume that at any given t these wavenumbers are applied instantaneously given the speed at which they are applied. The resulting data is a sequence of experimental absorption measures $\{\tilde{\mathcal{A}}_{ilt}\}_{\forall ilt}$. Figure 3.4 illustrates the eight experimental absorption measures for the full wavenumber range across all nine time points, and Figure 3.5 illustrates the average at each time.

Figure 3.4: Experimental Data

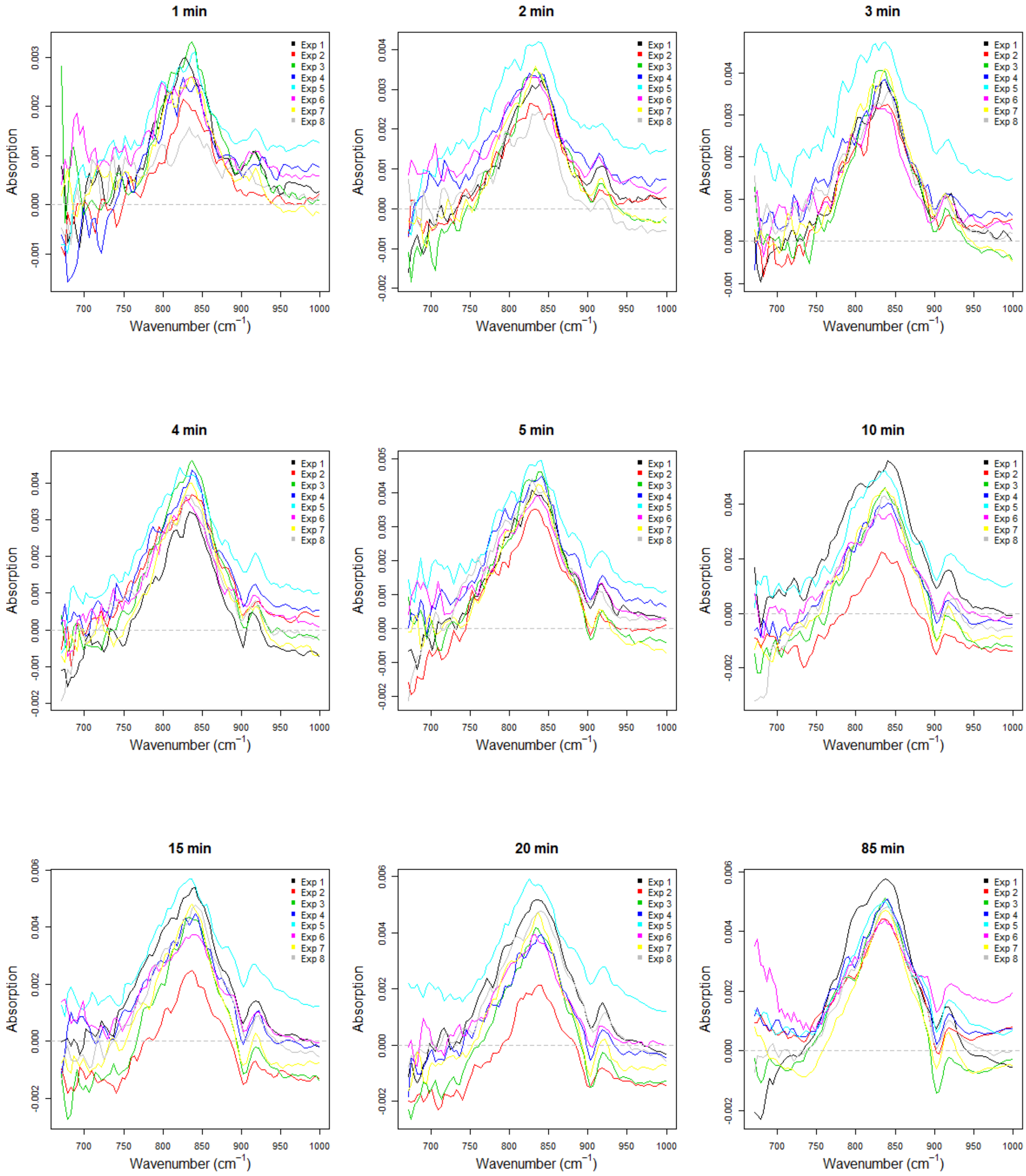
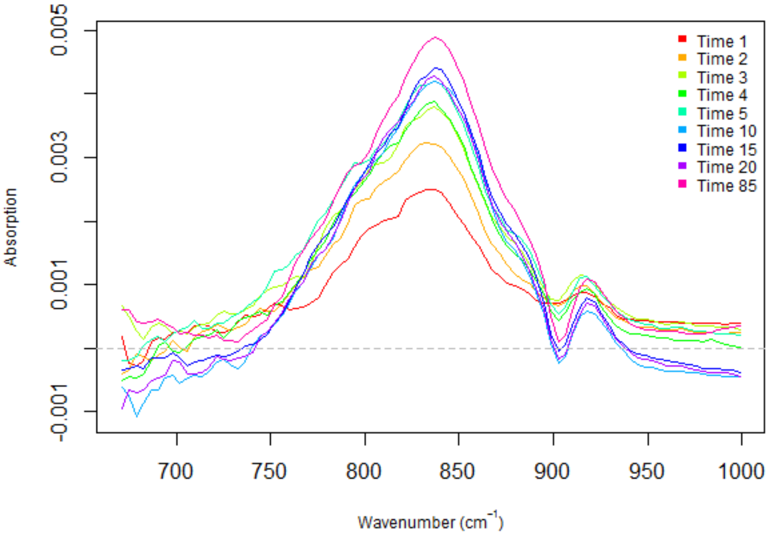


Figure 3.5: Averaged Experimental Data



The experimental absorption curves can be thought of as a snapshot of the system at one given point in time.

Chapter 4

Model

Recall the current MCR methodology which decomposes the absorption data into the following matrix-wise bilinear function

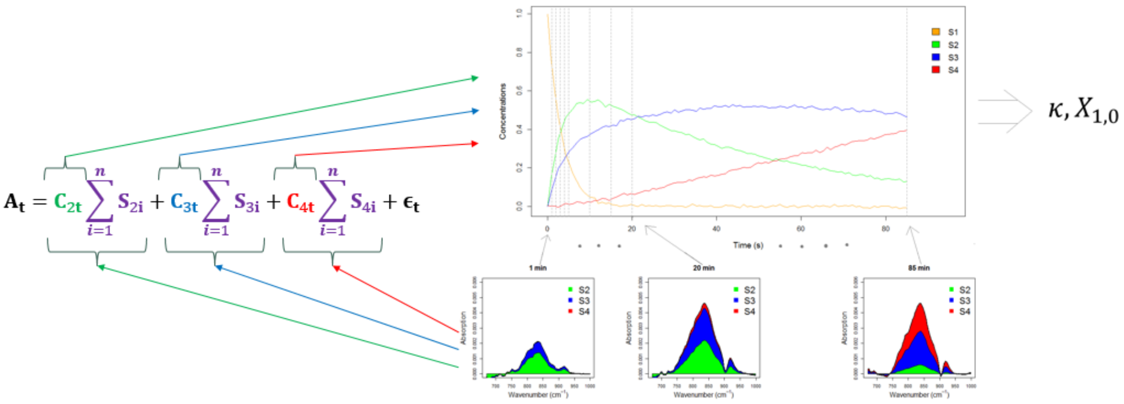
$$\tilde{\mathcal{A}} = \mathbf{C}\mathbf{S}^{\top} + \mathbf{E}$$

Alternatively, each absorption entry can be expressed as (k= number of species)

$$\tilde{\mathcal{A}}_{it} = \sum_{j=1}^{k} c_{tj} s_{ji} + \epsilon_{it} \qquad \forall i, \forall t$$

where $\mathbf{C}, \mathbf{S}^{\top}$ are estimated via alternating minimizations of some cost function, and error distribution assumptions for $\mathbf{E}$ are relaxed almost entirely. Figure 4.1 illustrates how the bilinear form establishes the link between the IR absorption process and the concentration profiles which in turn imply ODE parameters. The green, blue, red areas underneath the curves correspond to the *absorption* contributions of each species and the purple expressions represent the pure spectral components of each species at time t across all n wavenumbers.

Figure 4.1: Link between IR Absorption and individual concentrations


$A_t = C_{2t}\sum_{i=1}^{n} S_{2i} + C_{3t}\sum_{i=1}^{n} S_{3i} + C_{4t}\sum_{i=1}^{n} S_{4i} + \epsilon_t$
Concentrations
Time (s)
S1
S2
S3
S4
$\kappa, X_{1,0}$
1 min
20 min
85 min
Absorption
Wavenumber (cm^{-1})

It is worth noting however, that when the integration of the chemical ODE system can be expressed analytically, some MCR methods reduce the number of free parameters in the bilinear function by expressing $\mathbf{C}$ in closed form. For example, if the solution to the ODE can be expressed by some function $f : \mathbb{R} \mapsto \mathbb{R}^k$, then the concentration of all $j = 1, \ldots, k$ species is completely determined for all $t > 0$ and MCR formulates $\mathbf{C} \in \mathbb{R}^{T \times k}$ as [38]

$$\mathbf{C} = \begin{bmatrix} f_1(1;\kappa, X_0) & \ldots & f_k(1;\kappa, X_0) \\ \vdots & f_j(t;\kappa, X_0) & \vdots \\ f_1(T;\kappa, X_0) & \ldots & f_k(T;\kappa, X_0) \end{bmatrix}$$

However as mentioned in 1.2, no parametric forms are given to the pure spectral components $\mathbf{S}^\top \in \mathbb{R}^{k \times n}$.

We discuss a model which directly related the concentration profiles to the rate constants, κ. In addition, the model utilizes the theoretical data obtained from computational chemistry simulations (discussed in 3.2.2) to parameterize the pure spectral components of each species.

4.1 Infrared Absorption Model

Recall the set of 17 paired theoretical wavenumbers, μ, and relative absorptions, γ, corresponding to each species; $\{|S_2|, |S_3|, |S_4|\} = \{8, 5, 4\}$. Each pair corresponds to an IR wavenumber μ which, when applied to that bond, causes a significantly intense bond vibration (hence photon absorption); the absorption at that wavenumber is quantified by its respective theoretical absorption value γ.

Given this reaction structure between molecular bonds and the IR spectrum, for each theoretical component $k \in \{1, \ldots, |S_j|\}$ corresponding to species j considered separately, we would expect its true absorption to peak at that wavenumber μ_{jk} and diffuse at wavenumbers further away. For each theoretical component, we consider modeling the true absorption characteristic $\mathcal{A}_{jk}$ at any wavenumber ω, as a normalized Gaussian density function ϕ centered around that theoretical wavenumber μ_{jk}, with some scale parameter σ_{jk}: (we say normalized because $\int \phi(\omega)d\omega = 1$)

$$\mathcal{A}_{jk}(\omega) = \phi(\omega; \mu_{jk}, \sigma_{jk})$$
$$k \in \{1, 2, \ldots, |S_j|\,; j = 2, 3, 4\}$$

We generalize the absorption characteristic at wavenumber ω by species j from just one of its theoretical components, to all of its components $k = 1, 2, \ldots, |S_j|$. We take a linear combination of its $|S_j|$ density components; each component is weighted by its corresponding theoretical absorption intensity γ_{jk}. The absorption characteristic of only species j is modeled as a normalized mixture Gaussian density function of the form: (we say normalized because γ is normalized such that $\sum_k \gamma_{jk} = 1 \implies \int \mathcal{A}_j(\omega)d\omega = 1$)

$$\mathcal{A}_j(\omega) = \sum_{k=1}^{|S_j|} \gamma_{jk}\phi(\omega; \mu_{jk}, \sigma_{jk}) \qquad \forall j$$

Note that so far (i) our absorption function has not been time dependent; we always assume the theoretical data to be independent of time but rather dependent on the chemical characteristics of the bond and (ii) we have modeled the absorption with a normalized density mixture function which does not yet reflect the area underneath the curve.

We generalize further to model the true aggregate absorption at wavenumber ω at time t, specifying a linear combination of the three separate absorption characteristics of each species obtained above. Like the MCR, we weight each species spectral component (the mixture Gaussian density function $\mathcal{A}_j$), by a time dependent absorption contribution of species j at time t, α_{jt}. We obtain a linear combination of three mixture Gaussian density functions to describe the true aggregate absorption of the form:

$$\mathcal{A}_t(\omega) = \beta_t + \sum_{j=2}^{4} \alpha_{jt} \sum_{k=1}^{|S_j|} \gamma_{jk} \phi(\omega; \mu_{jk}, \sigma_{jk}) \qquad \forall t$$

Introducing $(\alpha_{2t}, \alpha_{3t}, \alpha_{4t})$ addresses (i) by making the function time dependent and also addresses (ii) by allowing the area underneath the absorption curve to be described by α_{jt}, thus obeying the Beer-Lambert Law. (Mathematically, $\int \mathcal{A}_t(\omega)d\omega = \alpha_{2t} + \alpha_{3t} + \alpha_{4t}$ for $\beta_t = 0$).

The unknown parameters defining the absorption function so far are related only to the IR absorption process; (β, σ, α). β_t is an intercept term that accounts for shifts along the absorption axis that are unrelated to the real absorption process, such as experimental error or recording error; such an intercept adds the flexibility of accounting for negative absorption readings in the data. σ_{jk} is the scale parameter of theoretical component k corresponding to bond j and α_{jt} is the absorption contribution of bond j at time t.

To establish the direct link between the IR absorptions, $\mathcal{A}_t(\omega)$, and concentration ODE parameters (κ, X_0), recall that α_{jt} is the absorption contribution of species j at time t as related to the IR absorption measurements. Alternatively, it can be interpreted as the relative concentration of species j at time t and hence an implied solution to the ODE system as defined in (3) and (4) where $X_{jt} \propto \alpha_{jt}$ (for convenience we refer to α as X, see Appendix C ii) for some given (κ, X_0). As such, to define an absorption function parameterized by $\Theta = (\kappa, X_0, \beta, \sigma)$, we require α_{jt} to be expressed in terms of (κ, X_0).

Both ODE systems defined in (3)-(4) are first order linear systems of differential equations; given reaction rate constants $\kappa = (\kappa_{12}, \kappa_{13}, \kappa_{23}, \kappa_{34})$ and initial concentrations $X_0 = (X_{1,0}, X_{2,0}, X_{3,0}, X_{4,0})$, the general solution $\{X_t = (X_{1t}, X_{2t}, X_{3t}, X_{4t})\}_{\forall t}$ to these systems can be expressed analytically by a function $f : \mathbb{R} \mapsto \mathbb{R}^4$ parameterized only by (κ, X_0) (see Appendix C i for derivation)

$$X_t = f(t; \kappa, X_0) = Qe^{\Lambda t}Q^{-1}X_0$$

where $\Lambda \in \mathbb{R}^{4\times4}$ is a diagonal matrix of eigenvalues of Ω ($\Lambda_{qq} = \lambda_q$,$\Lambda_{qp} = 0$, $\forall q \neq p$) and $Q \in \mathbb{R}^{4\times4}$ is a matrix of eigenvectors of Ω. Both Λ, Q are expressed just in terms of κ (see Appendix B for analytic forms).

Finally, we can express the true IR absorption function as parameterized by both the IR Spectroscopy parameters (β, σ) and concentration ODE parameters (κ, X_0), obtaining a direct link between the reaction rate constants and the IR absorption process which is of the form

$$\mathcal{A}_t(\omega) = \beta_t + \sum_{j=2}^{4} f_j(t; \kappa, X_0) \sum_{k=1}^{|S_j|} \gamma_{jk}\phi(\omega; \mu_{jk}, \sigma_{jk}) \qquad \forall t$$

In particular, given data of only a finite set of experimental wavenumbers, our true absorption model over the data becomes

$$\mathcal{A}_t(\omega_i) = \beta_t + \sum_{j=2}^{4} f_j(t; \kappa, X_0) \sum_{k=1}^{|S_j|} \gamma_{jk}\phi(\omega_i; \mu_{jk}, \sigma_{jk}) \qquad \forall i, \forall t$$

Note that α is no longer an explicit parameter in the absorption function as it has been redefined as $\alpha_t = f(t; \kappa, X_0)$ to establish the link.

4.2 Statistical Model of Measurement Error

As mentioned in section 2.2.1, the quantity reflecting the number of photons absorbed when experimental wavenumber ω_i is applied at time t is defined as a log-difference, $\tilde{\mathcal{A}}_{it} = \log\left(\frac{I_R}{I_S}\right)$.

We can interpret the transformed experimental absorptions, $\tilde{\mathcal{A}}_{it}$, as noisy observed quantities being generated from some process with mean $\mathcal{A}_t(\omega_i)$ and variance τ^2.

In particular, we consider the additive random measurement error model, $\tilde{\mathcal{A}}_{it} = \mathcal{A}_{it} + \epsilon_{it}$ (see Appendix F for justification), where

$$\epsilon_{it} \overset{iid}{\sim} \mathcal{N}ormal\left(0, \tau^2\right) \qquad \forall i, \forall t$$

Thus obtaining the Log-Likelihood function of the form

$$\ell(\Theta|\mathcal{D}) = -\frac{nT}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\tilde{\mathcal{A}}_{it} - \mathcal{A}_t(\omega_i; \beta, \sigma, \kappa, X_0)\right)^2$$

τ is a nuisance parameter which accounts for experimental errors and uncertainties, machine noise, as well as other sources of unexplained variances in the IR absorption process.

Chapter 5

Bayesian Inference

Due to the high dimensionality of the model parameters combined with variabilities in the data, both simulation studies and real data estimations would suggest not only a highly multi-modal likelihood surface, but also model sensitivity to different inputs. To increase the chances of our sampling algorithms exploring chemically plausible surface modes, the chemist provides sound beliefs regarding certain characteristics of the chemical system a priori which we incorporate into the model via carefully chosen prior density functions over Θ.

We consider a Bayesian model and specify a posterior distribution $p(\Theta|\mathcal{D}) \propto L(\Theta|\mathcal{D})\pi(\Theta)$ on the parameters $\Theta = (\kappa, X_0, \beta, \sigma, \tau)$ given data $\mathcal{D} = \{\omega_i, \tilde{A}_{it}\}_{\forall it}$.

5.1 Prior on κ

Plausible estimates for κ are provided based on repeated experimentation and other chemical knowledge:

$$\kappa_{12} : 0.05 \pm 0.01 \tag{5.1}$$

$$\kappa_{13} : 0.01 \pm 0.01 \tag{5.2}$$

$$\kappa_{23} : 0.01 \pm 0.01 \tag{5.3}$$

$$\kappa_{34} : 0.001 \pm 0.0005 \tag{5.4}$$

Although these reaction rate constants are unattainable and therefore not actually known, the chemist is nevertheless confident in their *proportional* relations to one another but less confident with the scaling at which these estimates have been previously obtained. As such, we scale the given rate constants (7)-(10) by $\zeta = 10$ (see Appendix D for derivation) which yields the following adjusted estimates a priori:

$$\kappa_{12} : 0.5 \pm 0.1$$

$$\kappa_{13} : 0.1 \pm 0.1$$

$$\kappa_{23} : 0.1 \pm 0.1$$

$$\kappa_{34} : 0.01 \pm 0.005$$

We take these as hyper parameters to model the uncertainty of κ under a joint Gaussian density function

$$h(\kappa) = \left(2\pi \left|\Sigma_\kappa\right|\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\left(\kappa - \mu_\kappa\right)^\top \Sigma_\kappa^{-1}\left(\kappa - \mu_\kappa\right)\right\}$$

where

$$\mu_\kappa = \begin{bmatrix} 0.5 \\ 0.1 \\ 0.1 \\ 0.01 \end{bmatrix}, \Sigma_\kappa = \begin{bmatrix} 0.1^2 & 0 & 0 & 0 \\ 0 & 0.1^2 & 0 & 0 \\ 0 & 0 & 0.1^2 & 0 \\ 0 & 0 & 0 & 0.005^2 \end{bmatrix}$$

5.2 Prior on α

Recall that $\alpha_t = f(t; \kappa, X_0) \in \mathbb{R}^4$ is the implied concentration of $\{S_1, S_2, S_3, S_4\}$ at time t; the implied solution to the concentration ODE system. It can be deduced from reactions (1) that in the long run, conditional on $\kappa > 0$, the chemical system will be dominated by S_4 with S_1, S_2, S_3 having been diminished to 0. Although the individual concentrations are experimentally unattainable at any time, it is strongly believed that there is still a presence of S_2 and S_3 at $t = 85min$. In particular, it is believed that the system is still dominated by S_2 at $t = 85min$ with $\alpha_{2,85} > \alpha_{3,85} > \alpha_{4,85}$ at approximately $70\% > 20\% > 10\%$ respectively (note that this assumption implies that S_1 has been largely diminished).

As α is not an explicit parameter in the model, we are unable to directly impose a prior density $\rho_\alpha(\alpha)$. However, α is expressed as the ODE solution, $\alpha_t = f(t; \kappa, X_0)$. As such, we can impose a prior density over the proportions of $f_2(85; \kappa, X_0), f_3(85; \kappa, X_0), f_4(85; \kappa, X_0)$, namely, $\rho_\alpha(\kappa)$. To reflect these proportions, we impose a joint Gaussian-like density over these proportions at $t = 85min$ of the form

$$\begin{aligned}\rho_\alpha(\kappa) &= \left(2\pi\,|\Sigma_{\alpha^*}|\,\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\left(\frac{f^*(\kappa, X_0)}{\|f^*(\kappa, X_0)\|_1} - \mu_{\alpha^*}\right)^\top \Sigma_{\alpha^*}^{-1}\left(\frac{f^*(\kappa, X_0)}{\|f^*(\kappa, X_0)\|_1} - \mu_{\alpha^*}\right)\right\}\left|\frac{\partial f^*(\kappa, X_0)}{\partial \kappa}\right| \\ &= g(\kappa)\left|\frac{\partial f^*(\kappa, X_0)}{\partial \kappa}\right|\end{aligned}$$

where

$$f^*(\kappa, X_0) = \begin{bmatrix} f_2(85; \kappa, X_0) \\ f_3(85; \kappa, X_0) \\ f_4(85; \kappa, X_0) \end{bmatrix}, \alpha^* = \begin{bmatrix} \alpha_{2,85} \\ \alpha_{3,85} \\ \alpha_{4,85} \end{bmatrix} \mu_{\alpha^*} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix}, \Sigma_{\alpha^*} = \begin{bmatrix} \sigma_{\alpha_2}^2 & 0 & 0 \\ 0 & \sigma_{\alpha_3}^2 & 0 \\ 0 & 0 & \sigma_{\alpha_4}^2 \end{bmatrix}$$

Σ_{α^*} is taken as a tuning hyper parameter and $\left|\frac{\partial f^*(\kappa, X_0)}{\partial \kappa}\right|$ is the determinant of the Jacobian matrix when applying the change of variables $\alpha^* \to \kappa$.

It is unclear, however, what the structure of such a Jacobian may be given $\alpha^* \to \kappa \implies \mathbb{R}^3 \to \mathbb{R}^4$. We instead consider a prior density on the κ parameterized concentrations at $t = 85min$ of the form

$$\rho_\alpha(\kappa) \sim g(\kappa)$$

In effect, this density decreases the likelihood over κ regions that predict concentrations at $t = 85min$ deviating far from the relative proportions of $S_2(70\%) > S_3(20\%) > S_4(10\%)$ and increases the likelihood over κ regions that predict otherwise.

Combining the density $h(\kappa)$ defined in 5.1 with $g(\kappa)$ above, we obtain a final prior density over κ

$$\pi(\kappa) \sim g(\kappa) \cdot h(\kappa)$$

5.3 Prior on σ

Recall that the model is a mixture of three mixture Gaussian density functions, each centered around a theoretical frequency μ_{jk} with a scale of σ_{jk}; a total of 17 scale parameters must be estimated from the data. When estimating $\sigma \in \mathbb{R}^{17}$ to fit low dimensional data $(\omega_i, \tilde{A}_{it}) \in \mathbb{R}^2$, we might expect there to be many combinations of $\sigma = (\sigma_1, \ldots, \sigma_{17})$ which are very far from eachother in $\mathbb{R}^{17}$ (in the Euclidean sense) yet all provide very reasonable fits to the IR absorption curves; in turn we expect a multi-modal likelihood surface. Indeed, repeated simulations would show vastly different combinations of σ to generate indistinguishable data. Moreover, the chemically implausible outputs of $\sigma_{jk} \to \infty$ would frequently occur under estimation.

However, if the chemical simulations suggest that significant absorptions would be observed at and around the neighborhood of each theoretical wavenumber μ_{jk}, we would intuitively expect their corresponding densities to peak in these areas. In order to explore reasonable σ

regions, we specify a light tailed prior density function, $\pi(\sigma_{jk}) \overset{ind}{\sim} \mathcal{W}eibull(\delta_{jk}, \xi_{jk})$, over σ

$$\pi(\sigma_{jk}) = \frac{\delta_{jk}}{\xi_{jk}}\Big(\frac{\sigma_{jk}}{\xi_{jk}}\Big)^{\delta_{jk}-1} \exp\Big\{-\Big(\frac{\sigma_{jk}}{\xi_{jk}}\Big)^{\delta_{jk}}\Big\} \qquad \forall j, \forall k$$

where $\delta_{jk}, \xi_{jk} \in \mathbb{R}^+$ are the shape and scale hyper parameters respectively which we take as tuning parameters. Note that the Weibull distribution is light tailed for $\delta_{jk} > 1$ which we impose in order to decrease the likelihood at implausibly high values of σ a priori.

5.4 Other Priors

Referring to the aggregate process $\{Y_{At}\}$ in Figure 3.1, we see that the aggregate concentration has begun to level off; in fact the chemist strongly believes the curve should be theoretically flat after $t = 20min$. Defining $X_{At} = X_{2t} + X_{3t} + X_{4t}$ as the aggregate concentration, it can be shown that $\lim\limits_{t\to\infty} X_{At} \to X_{1,0}$ (see Appendix E). As such the initial concentration of S_1, $X_{1,0}$, can be estimated non-parametrically as the total area underneath the absorption curve at the greatest time point, $t = 85min$. Namely, we compute the Riemann sum of the absorption curve at $t = 85min$

$$\hat{X}_{1,0} = \sum_{i=1}^{n} (\omega_{i+1} - \omega_i)\tilde{\mathcal{A}}_{i,85}$$

and model the uncertainty of the initial concentration under a Gaussian density

$$\pi(X_{1,0}) = \frac{1}{\sqrt{2\pi}\sigma_{X_{1,0}}} \exp\Big\{-\frac{1}{2}\Big(\frac{X_{1,0} - \hat{X}_{1,0}}{\sigma_{X_{1,0}}}\Big)^2\Big\}$$

where $\sigma_{X_{1,0}}$ is taken as a tuning standard deviation hyper parameter.

The parameters (β, τ) account for experimental errors and shifts in the data which we assume are unexplained, thus we assume a flat prior density $\pi(\beta, \tau) \propto 1$.

Considering a jointly independent parameter set Θ, we combine the Likelihood function with the joint prior distribution, obtaining a Posterior distribution of the following log form

$$\begin{aligned}\log p(\Theta|\mathcal{D}) =& -\frac{nT}{2}\log(\tau^2) - \frac{1}{2\tau^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big(\tilde{\mathcal{A}}_{it} - \mathcal{A}_t(\omega_i;\beta,\sigma,\kappa,X_0)\Big)^2 \\ & - \frac{1}{2}\left(\frac{f^*(\kappa,X_0)}{\|f^*(\kappa,X_0)\|_1} - \mu_{\alpha^*}\right)^\top \Sigma_{\alpha^*}^{-1}\left(\frac{f^*(\kappa,X_0)}{\|f^*(\kappa,X_0)\|_1} - \mu_{\alpha^*}\right) \\ & - \frac{1}{2}\Big(\kappa - \mu_\kappa\Big)^\top \Sigma_\kappa^{-1}\Big(\kappa - \mu_\kappa\Big) \\ & + \sum_{j=2}^{4}\sum_{k=1}^{|S_j|}(\delta_{jk} - 1)\log\sigma_{jk} - \sum_{j=2}^{4}\sum_{k=1}^{|S_j|}\Big(\frac{\sigma_{jk}}{\xi_{jk}}\Big)^{\delta_{jk}} \\ & - \frac{1}{2}\Big(\frac{X_{1,0} - \hat{X}_{1,0}}{\sigma_{X_{1,0}}}\Big)^2 + c\end{aligned}$$

where c is a constant term free of Θ.

Chapter 6

Simulation

We simulate plausible concentration curves under the 3 reaction and 4 reaction systems such that their trends are consistent with what the prior belief: $X_{2,85} > X_{3,85} > X_{4,85}$. The simulated process is shown in the first two figures, followed by histograms of the parameter posterior samples which are estimated from the simulated IR Absorption curves under a flat joint prior distribution $\pi(\kappa, X_0) \propto 1$.

6.1 4 Reaction System

We simulate IR Absorption curves under the following chemically plausible parameter values:

$$\kappa = (0.4, 0.2, 0.001, 0.005)$$

$$X_0 = (0.4, 0, 0, 0)$$

$$\sigma = (24, 22, 25, 25, 11, 9, 25, 17, 23, 25, 13, 8, 13, 16, 15, 15, 16)$$

$$\beta = (0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\tau = 0.00005$$

The Concentration curves and IR Absorption curves corresponding to the above parameters are illustrated in Figure 6.1 and Figure 6.2 respectively.

Figure 6.1: Simulated Concentration Curves

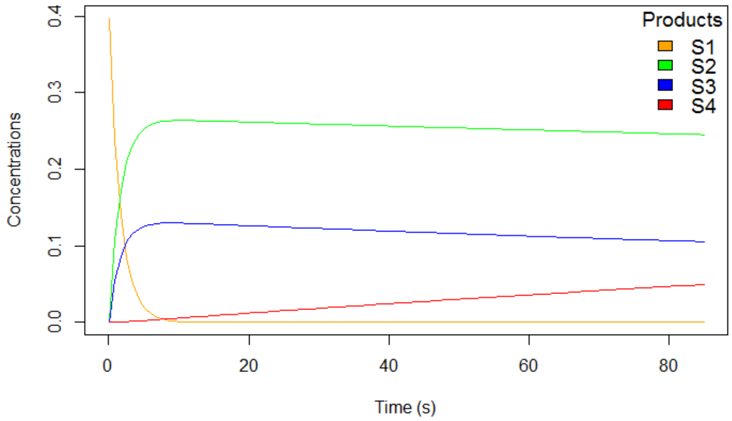
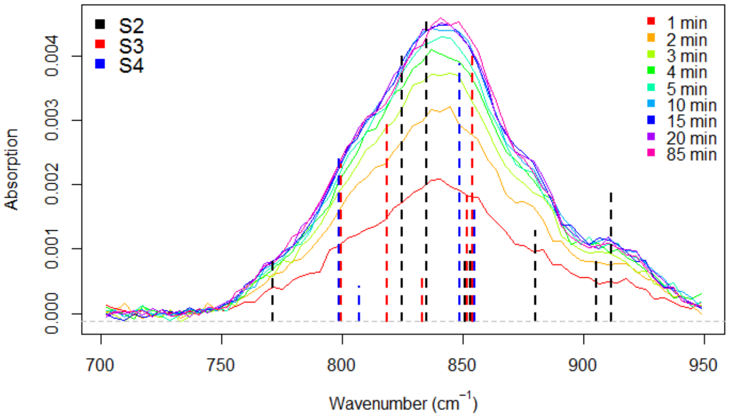


Figure 6.2: Simulated IR Absorption Curves



We re-estimate $(\kappa, X_{1,0}, \sigma)$ under fixed values of $(\hat{\beta}, \hat{\tau})$, in particular, assuming the curves are well positioned along the Absorption axis and assuming a known noise variance τ^2:

$$\hat{\beta} = (0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\hat{\tau} = 0.00005$$

Further, we set $[\Sigma_{\alpha^*}]_{jj} = [\Sigma_\kappa]_{jj} = \sigma_{X_{1,0}} = 1,000,000$ which implies flat priors over the ODE parameters, $\pi(\kappa, X_0) \propto 1$. The flat prior is chosen in order to examine how well the model can estimate the parameters of interest with heavier reliance on the data rather than specified knowledge a priori.

However, given the high dimensionality of σ, we control for a chemically plausible range by specifying a light tailed Weibull distribution such that $Quantile_\sigma(99.99\%) \approx 30$

$$\pi(\sigma_{jk}) \overset{ind}{\sim} \mathcal{W}eibull(\delta_{jk} = 2, \xi_{jk} = 10) \qquad \forall j, \forall k$$

We obtain the following posterior samples of $(\kappa, X_{1,0}, \sigma)$ from $p(\kappa, X_{1,0}, \sigma | \hat{\beta}, \hat{\tau}, \mathcal{D})$ after 5,000 sampling iterations

Figure 6.3: Posterior Estimates of κ,X_0

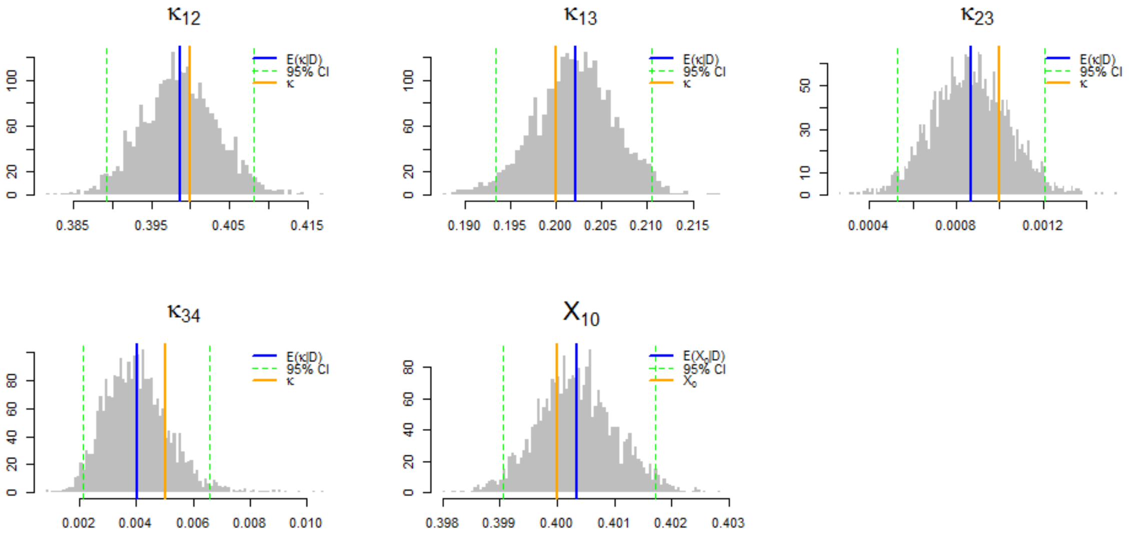
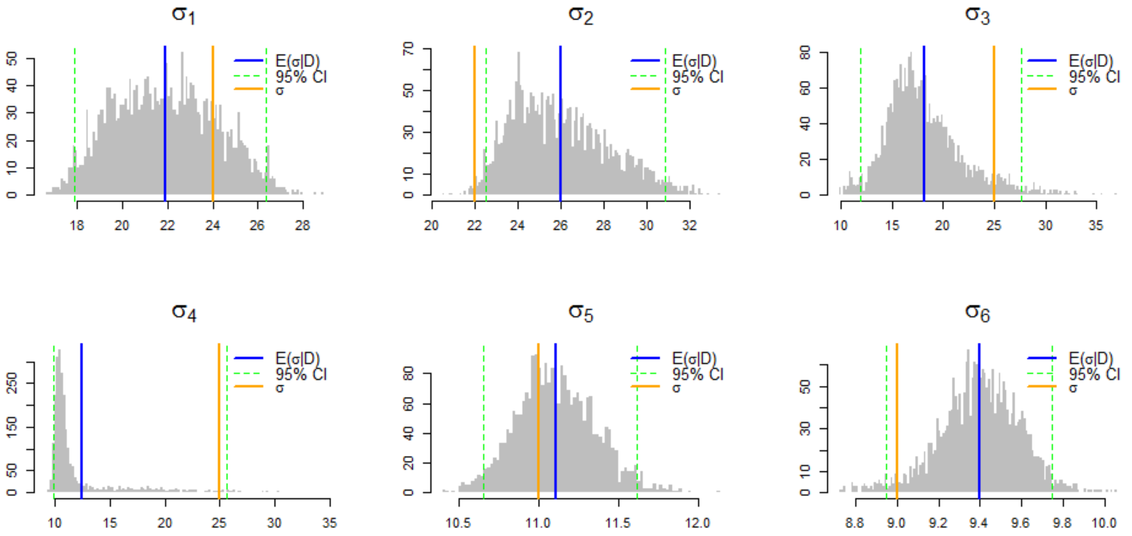
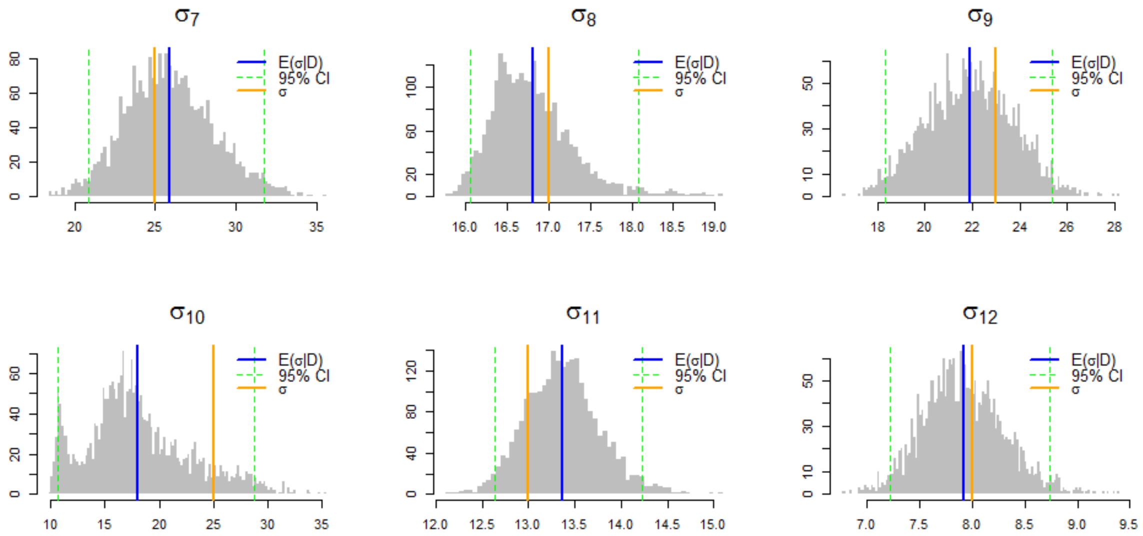
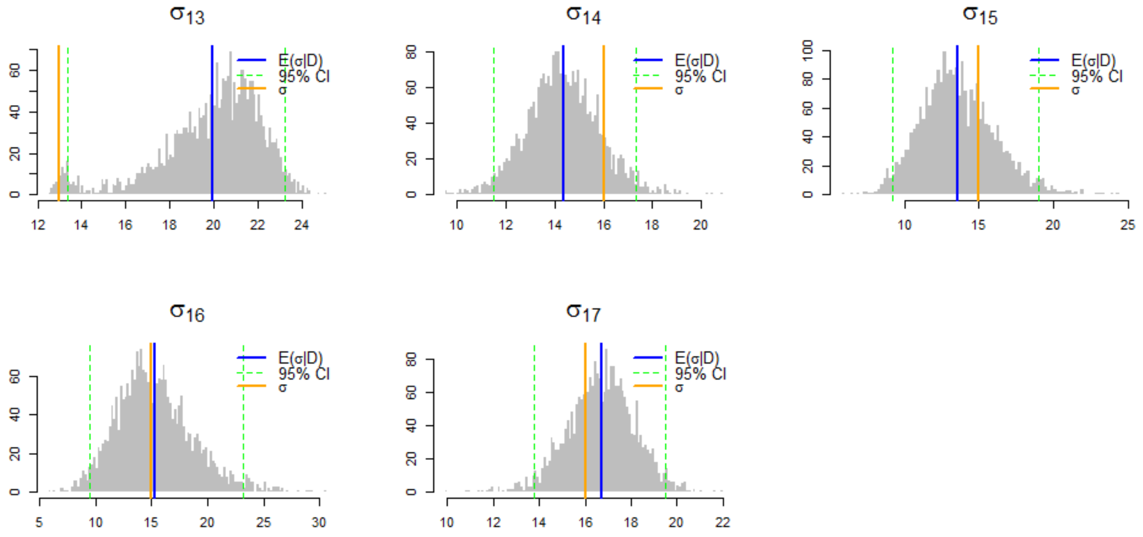


Figure 6.4: Posterior Estimates of σ

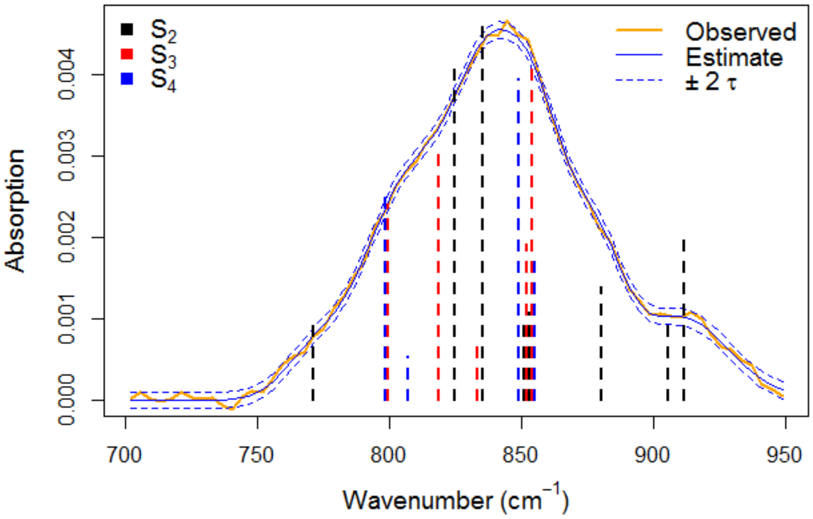



σ_7
σ_8
σ_9
σ_{10}
σ_{11}
σ_{12}
E(σ|D)
95% CI
σ
0 20 40 60 80
20 25 30 35
0 20 60 100
16.0 16.5 17.0 17.5 18.0 18.5 19.0
0 10 30 50
18 20 22 24 26 28
0 20 40 60
10 15 20 25 30 35
0 40 80 120
12.0 12.5 13.0 13.5 14.0 14.5 15.0
0 10 30 50
7.0 7.5 8.0 8.5 9.0 9.5


σ_{13}
σ_{14}
σ_{15}
σ_{16}
σ_{17}
E(σ|D)
95% CI
σ
0 20 40 60
12 14 16 18 20 22 24
0 20 40 60 80
10 12 14 16 18 20
0 20 40 60 80 100
10 15 20 25
0 20 40 60
5 10 15 20 25 30
0 20 40 60 80
10 12 14 16 18 20 22

Inspecting the above posterior histograms, we can see that almost all parameters were recovered within the 95% Bayesian credible interval. Moreover, the estimates were obtained using no prior information on κ, X_0. Figure 6.5 shows the resulting curve estimate at one time point, $t = 85min$.

Figure 6.5: Estimated Curve at t=85min



6.2 3 Reaction System

We simulate IR Absorption curves under the following chemically plausible parameter values:

$$\kappa = (0.5, 0.005, 0.015)$$

$$X_0 = (0.4, 0, 0, 0)$$

$$\sigma = (28, 28, 20, 20, 15, 30, 30, 15, 25, 15, 10, 7, 19, 20, 22, 17, 26)$$

$$\beta = (0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\tau = 0.00005$$

The Concentration curves and IR Absorption curves corresponding to the above parameters are illustrated in Figure 6.6 and Figure 6.7 respectively.

Figure 6.6: Simulated Concentration Curves

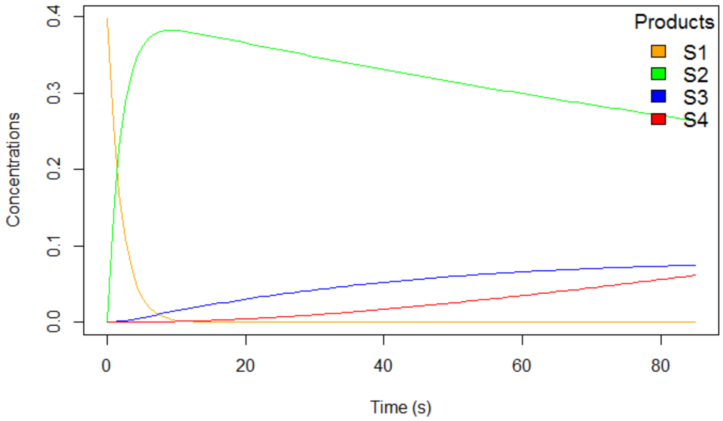
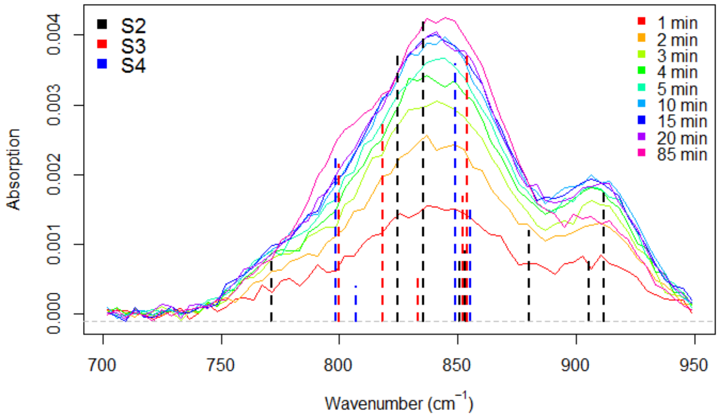


Figure 6.7: Simulated IR Absorption Curves



We re-estimate $(\kappa, X_{1,0}, \sigma)$ under fixed values of $(\hat{\beta}, \hat{\tau})$, in particular, assuming the curves are well positioned along the Absorption axis and assuming the correct noise variance τ^2:

$$\hat{\beta} = (0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\hat{\tau} = 0.00005$$

Further, we set $[\Sigma_{\alpha^*}]_{jj} = [\Sigma_\kappa]_{jj} = \sigma_{X_{1,0}} = 1,000,000$ which implies flat priors over the ODE parameters, $\pi(\kappa, X_0) \propto 1$. The flat prior is chosen in order to examine how well the model can estimate the parameters of interest with heavier reliance on the data rather than specified knowledge a priori.

However, given the high dimensionality of σ, we control for a chemically plausible range by specifying the following prior distribution:

$$\pi(\sigma_{jk}) \overset{ind}{\sim} \mathcal{W}eibull(\delta_{jk} = 2, \xi_{jk} = 10) \qquad \forall j, \forall k$$

We obtain the following posterior samples of $(\kappa, X_{1,0}, \sigma)$ from $p(\kappa, X_{1,0}, \sigma | \hat{\beta}, \hat{\tau}, \mathcal{D})$ after 5,000 sampling iterations

Figure 6.8: Posterior Estimates of κ,X_0

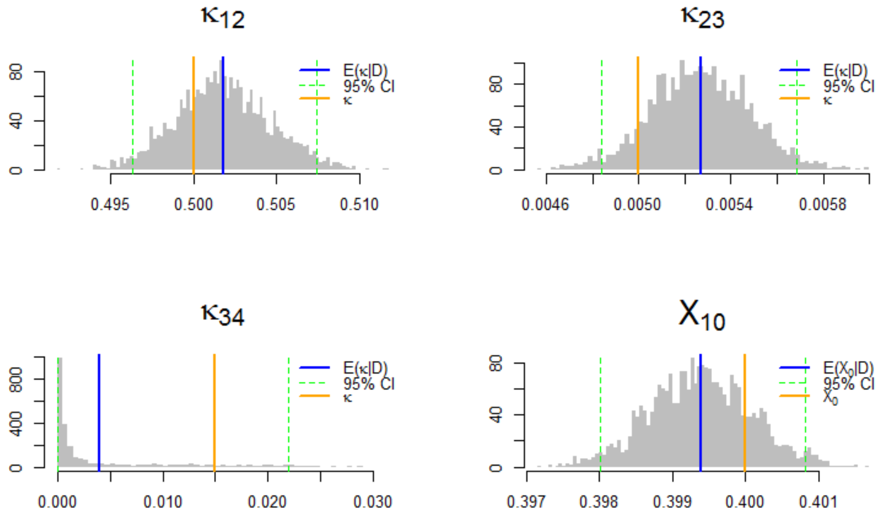
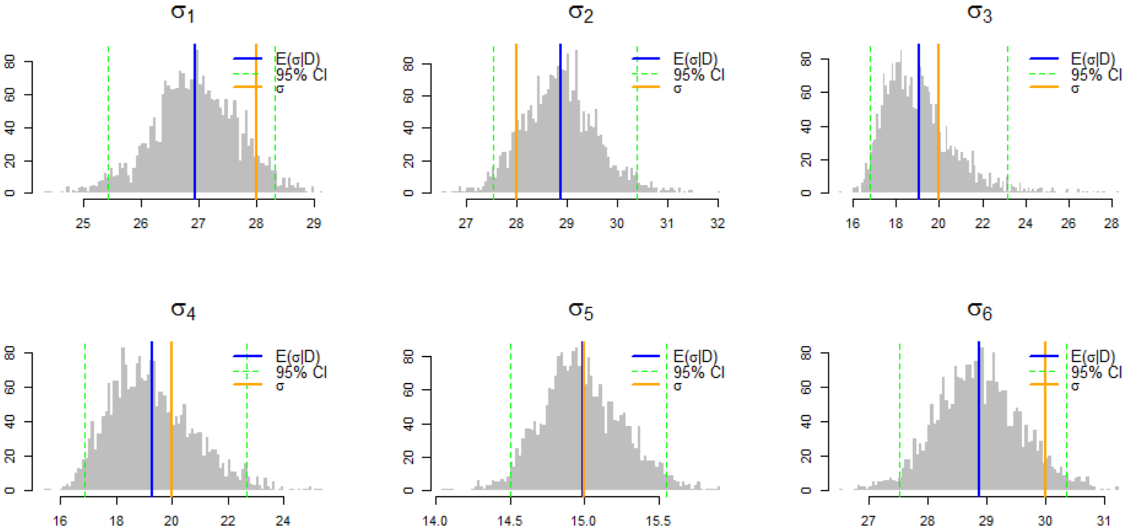
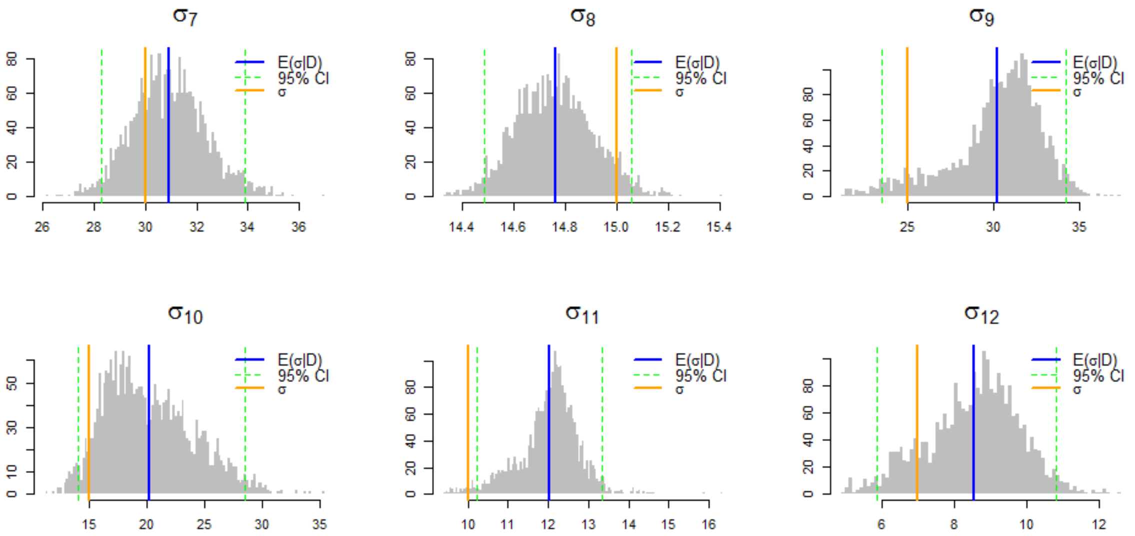
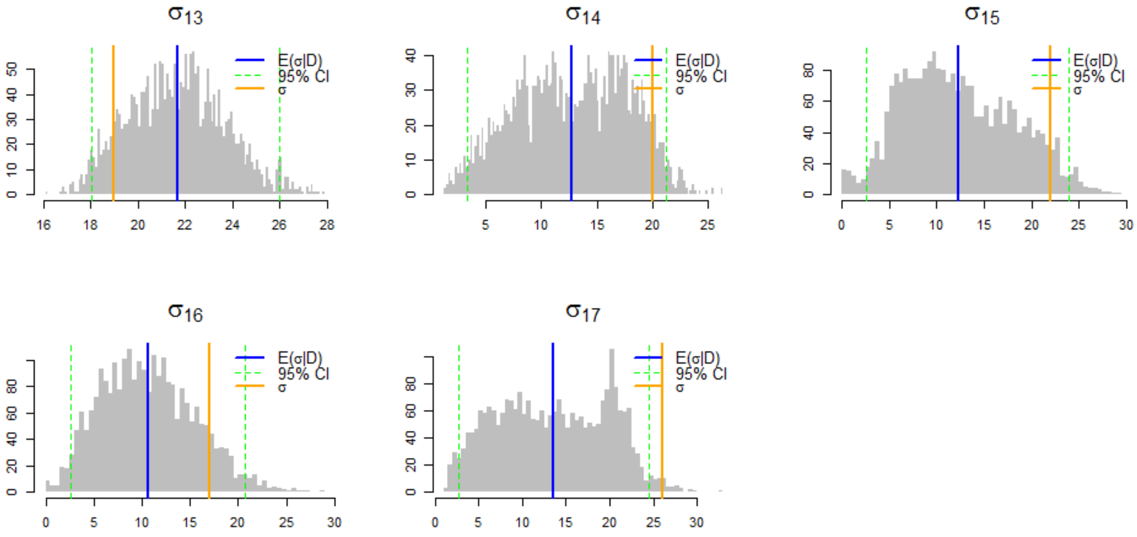


Figure 6.9: Posterior Estimates of σ





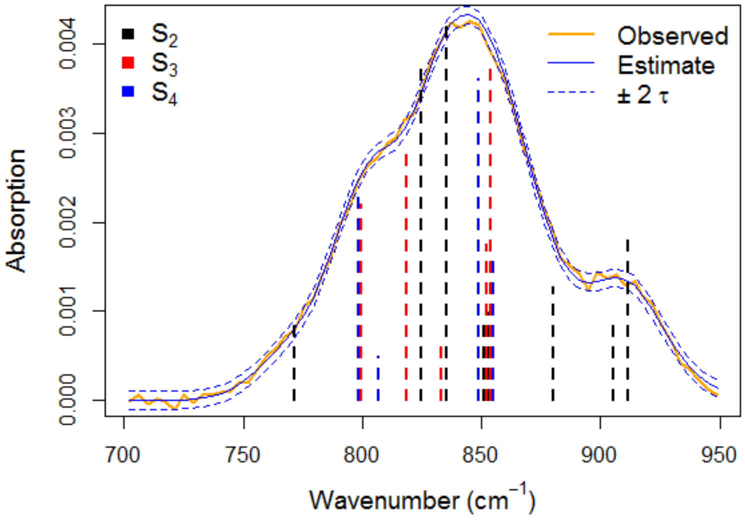
σ_7
0
20
40
60
80
26
28
30
32
34
36
E(σ|D)
95% CI
σ
σ_8
0
20
40
60
80
14.4
14.6
14.8
15.0
15.2
15.4
E(σ|D)
95% CI
σ
σ_9
0
20
40
60
80
25
30
35
E(σ|D)
95% CI
σ
σ_{10}
0
10
30
50
15
20
25
30
35
E(σ|D)
95% CI
σ
σ_{11}
0
20
40
60
80
10
11
12
13
14
15
16
E(σ|D)
95% CI
σ
σ_{12}
0
20
40
60
80
6
8
10
12
E(σ|D)
95% CI
σ



σ_{13}
0
10
20
30
40
50
16
18
20
22
24
26
28
E(σ|D)
95% CI
σ
σ_{14}
0
10
20
30
40
5
10
15
20
25
E(σ|D)
95% CI
σ
σ_{15}
0
20
40
60
80
0
5
10
15
20
25
30
E(σ|D)
95% CI
σ
σ_{16}
0
20
40
60
80
0
5
10
15
20
25
30
E(σ|D)
95% CI
σ
σ_{17}
0
20
40
60
80
0
5
10
15
20
25
30
E(σ|D)
95% CI
σ

Inspecting the above posterior histograms, we can see that all parameters were recovered within the 95% Bayesian credible interval. Moreover, the estimates were obtained using no prior information on κ, X_0. Further, we are able to obtain very good fit to the IR Absorption curves as seen in Figure 6.10.

Figure 6.10: Estimated Curve at $t = 85min$



6.3 Sensitivity Analysis

Here we show indistinguishable fits to the data under different parameters estimates. In particular, we focus on the vastly different σ estimates that can be obtained under a less informative prior and the consequence such a prior will have on the ODE parameters. The parameters are re-estimated with the same inputs as for their respective simulations above, but under a less informative prior on σ. We impose a prior density with $Quantile_{\sigma}(99.99\%) \approx$

2250 of the form

$$\pi(\sigma_{jk}) \overset{ind}{\sim} \mathcal{L}og - \mathcal{N}ormal(\mu_{\sigma_{jk}} = 4, \sigma_{\sigma jk} = 1) \qquad \forall j, \forall k$$

6.3.1 4 Reaction System

Figure 6.11: Posterior Estimates of κ,X_0

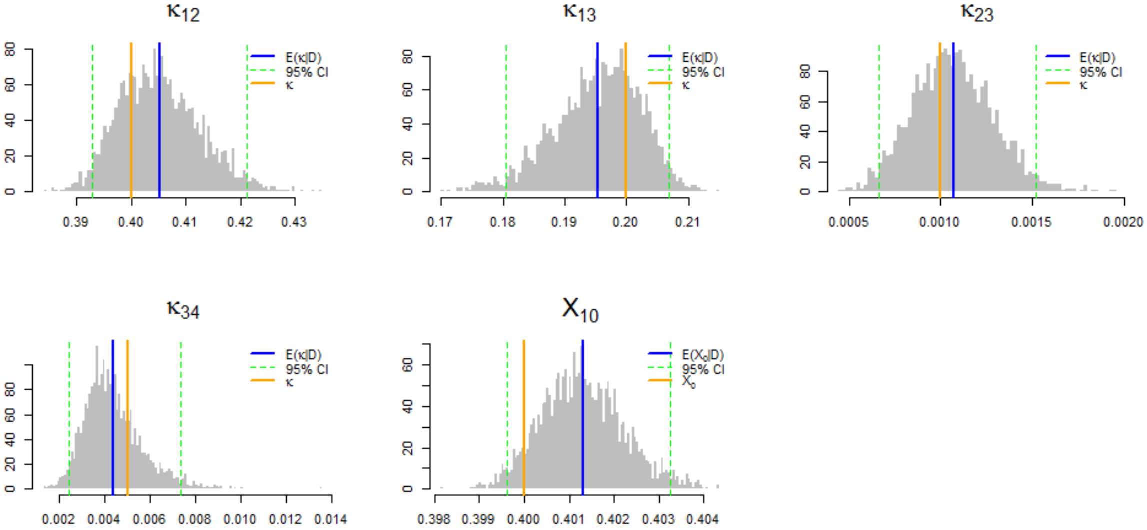
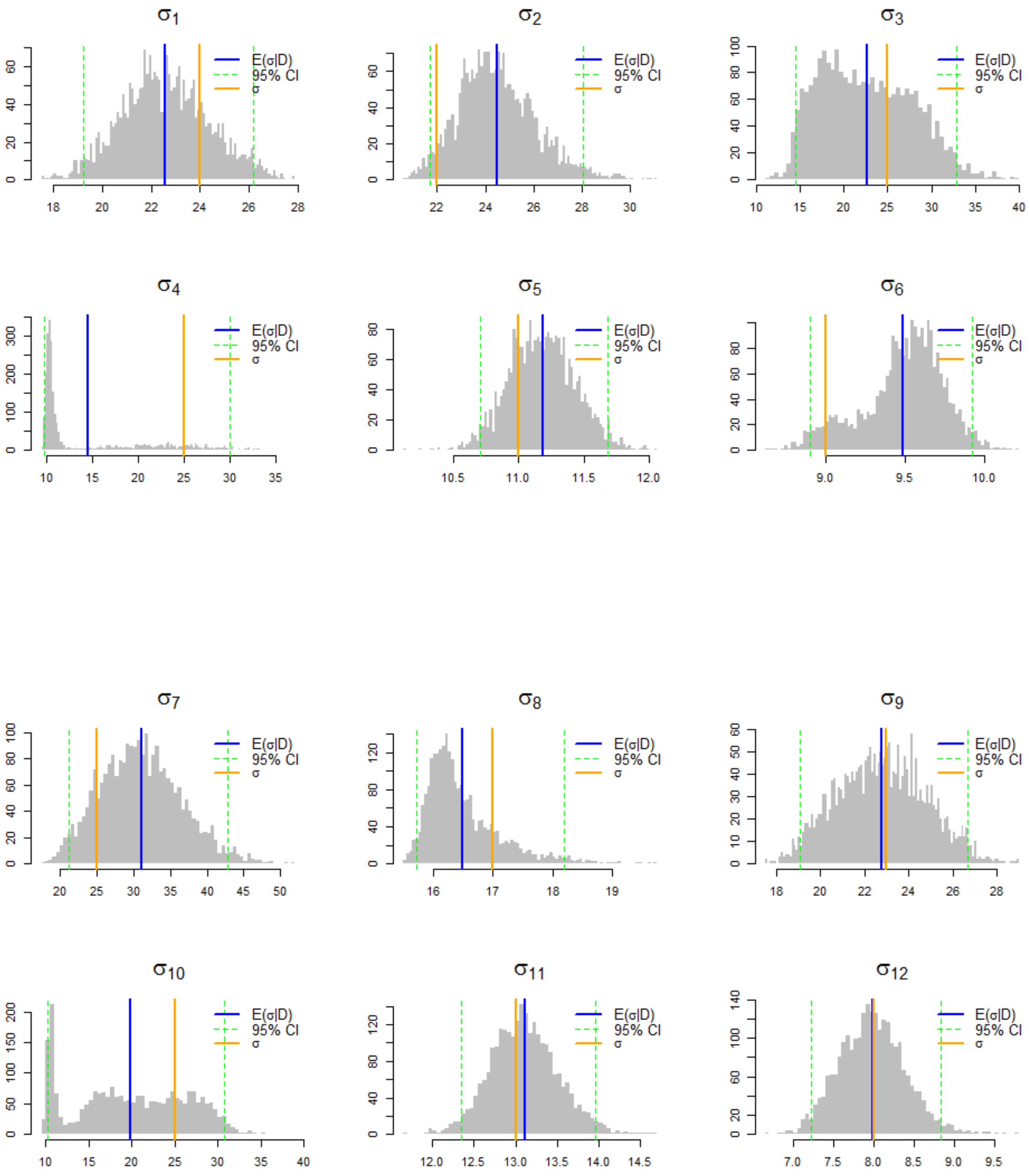


Figure 6.12: Posterior Estimates of σ



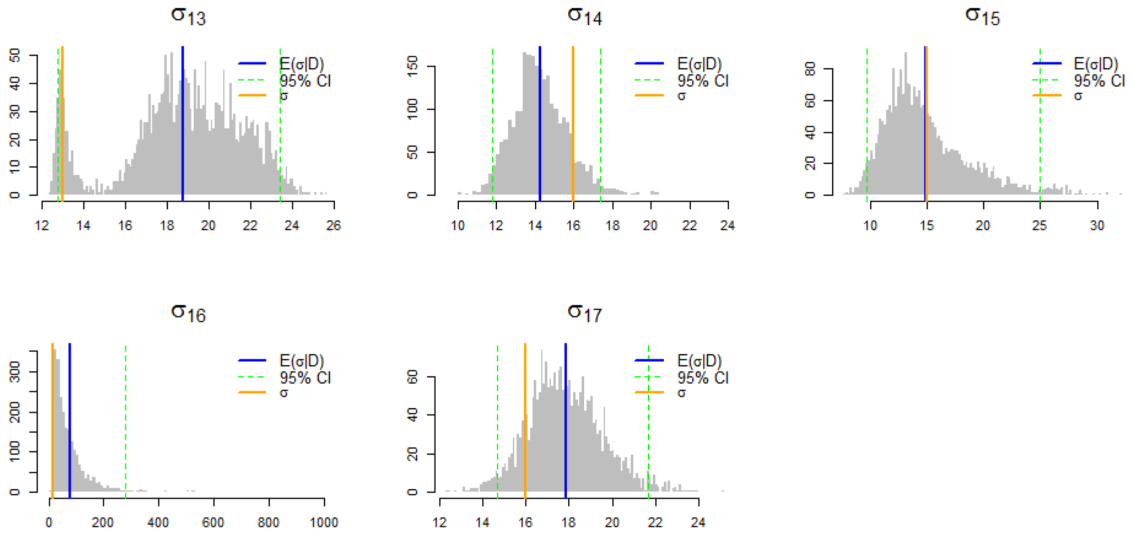
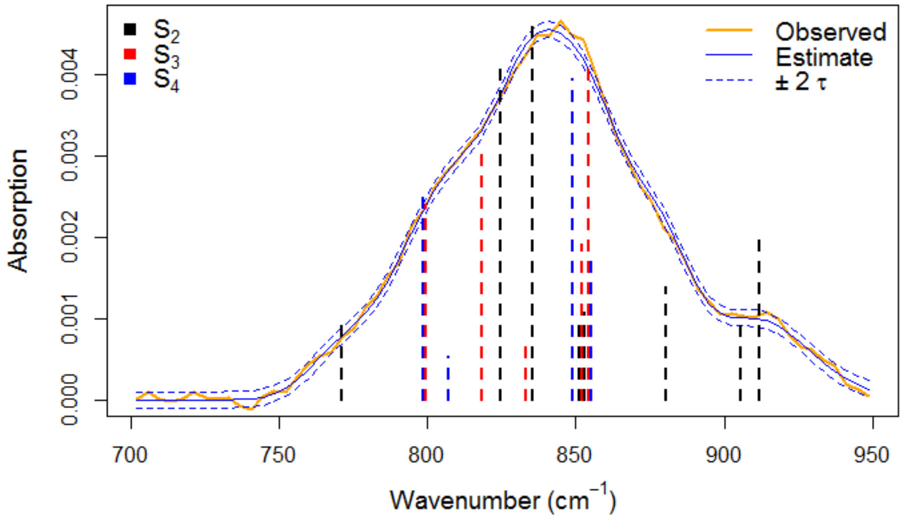


Figure 6.13: Estimated Curve at 85 min



Examining the marginalized posterior density histograms, it can be seen that most parameters have been recovered within the 95% Bayesian credible intervals.

However, it is interesting to note that some of the histograms exhibit multi modality; namely $\sigma_3, \sigma_4, \sigma_6, \sigma_{10}, \sigma_{13}$. In particular, examining the histograms of $\sigma_4, \sigma_6, \sigma_{13}$, we see their true values from simulation to be centered at the less dense modes. This might indicate that the simulated data is providing some evidence for plausibility of the regions around the true values, however the evidence seems to be very weak given the significantly lower density at those regions. This may suggest that even under simulated data, the model can be very sensitive with respect to σ as there may exist a large number of plausible combinations resulting in good fits to the data.

6.3.2 3 Reaction System

Figure 6.14: Posterior Estimates of κ,X_0

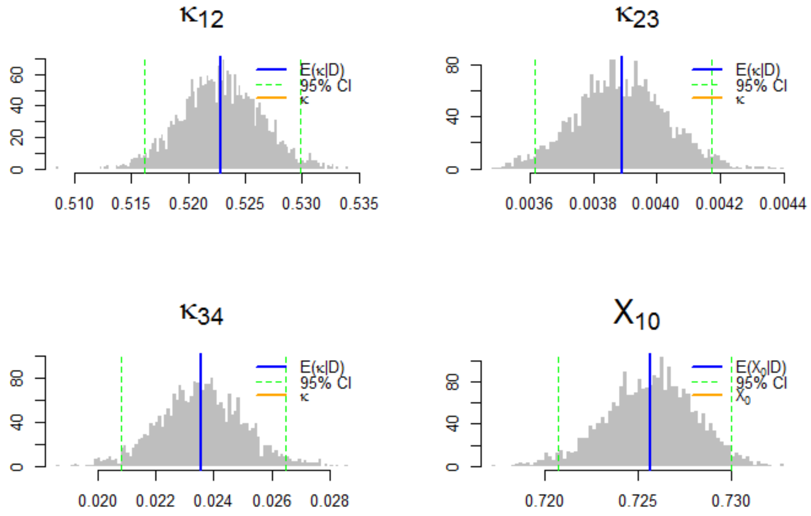
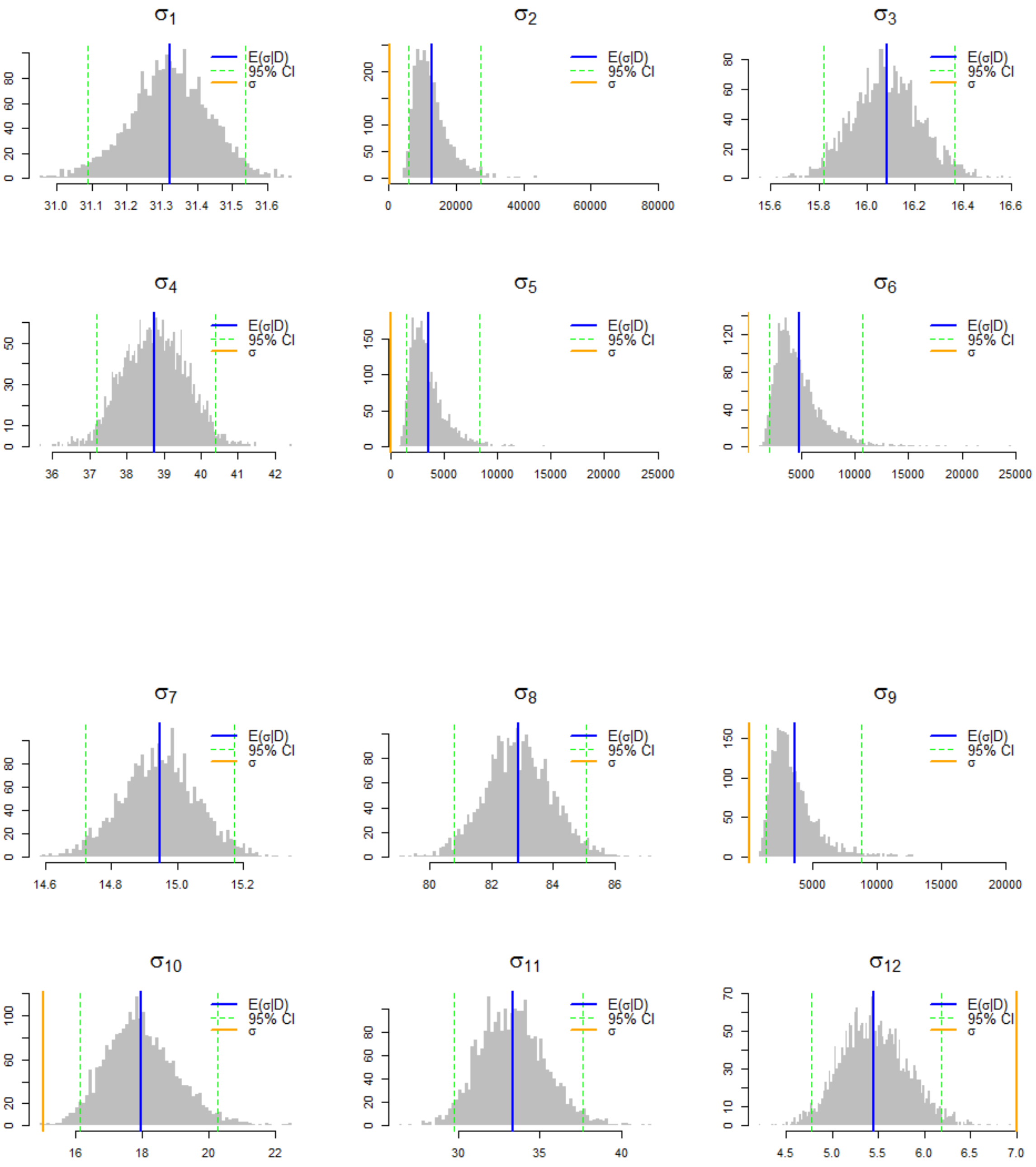
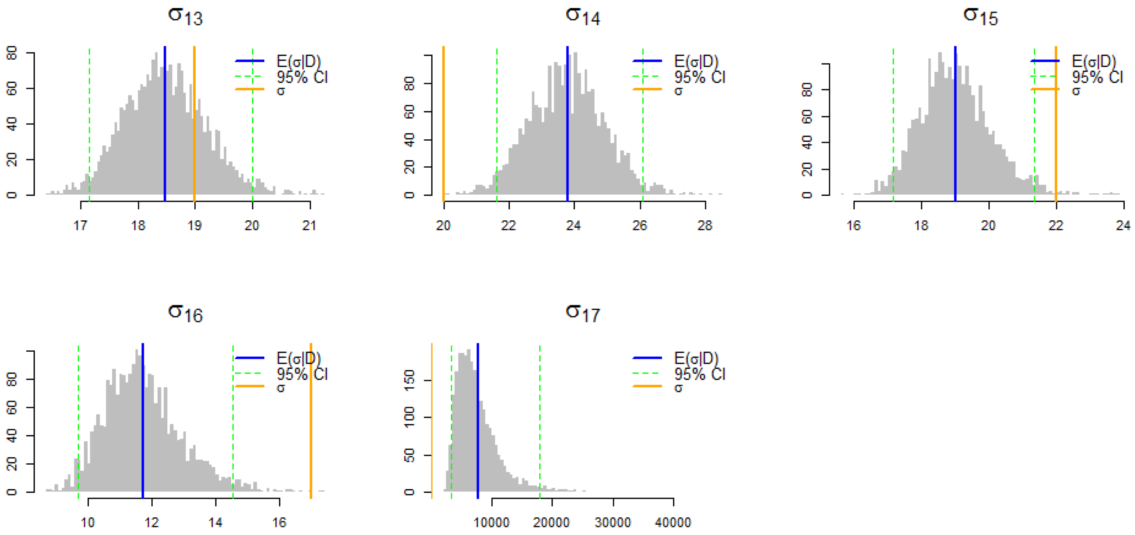


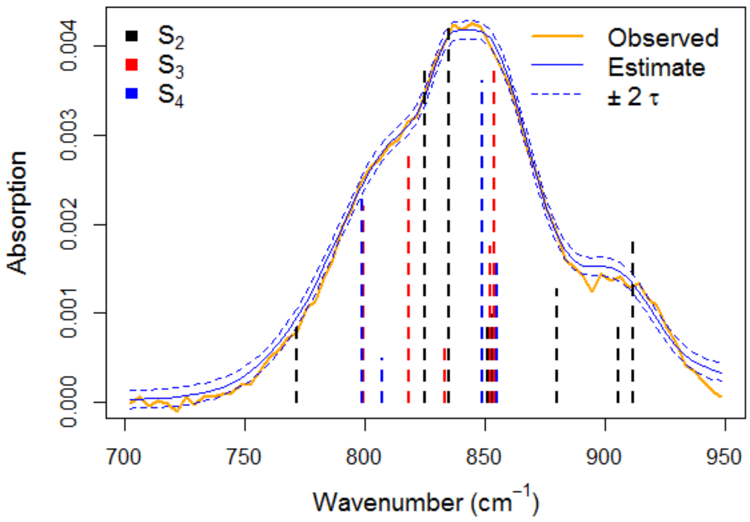
Figure 6.15: Posterior Estimates of σ





Here we see that under a less informative prior on σ, we obtain vastly different estimates of both κ and σ; the true values are too far from the posterior estimates to be seen on the histograms. Moreover, many σ estimates are too large to be considered plausible. Figure 6.16 shows the corresponding estimated IR Absorption curve at one chosen time point, $t = 85min$.

Figure 6.16: Estimated Curve at 85 min



We can see that under vastly different and implausible parameter values, we still obtain very good fits to the data. Moreover, these fits are indistinguishable from the fits obtained under the more informative Weibull prior density.

Chapter 7

Application

7.1 Experimental Data Processing

As previously mentioned, IR Absorption is measured after reading photon intensity through a sample and reference source. Due to the chemical properties of a given mixture, it will absorb significant amounts of photons at certain wavenumber ranges (absorbing ranges) and little to no photons at other wavenumber ranges (non-absorbing ranges). Naturally, a significantly greater amount of photons is expected to be absorbed by the sample cell than the reference cell within absorbing ranges, and an even amount by both cells within non-absorbing ranges; thus expecting absorption readings of 0 within non-absorbing ranges. The absorptions observed in these ranges are referred to as baseline absorptions; we have verified that the non-absorbing ranges include all $\omega < 700$ and $\omega > 950$.

It is evident from Figure 3.4 that the locations of each experiment along the vertical axis are noticeably away from 0 at non-absorbing ranges; it has been verified that these shifts are attributed to experimental errors. To correct for this, we center the experimental curves at each time by subtracting their mean experimental absorption readings for $\omega > 950$, thus approximately bringing the baseline to 0 as shown for $t = 85min$ in Figure 7.1. Figure 7.2 shows the average of all eight experimental curves at each time point after correcting for baseline misalignment.

Figure 7.1: Baseline Adjustment for t=85min

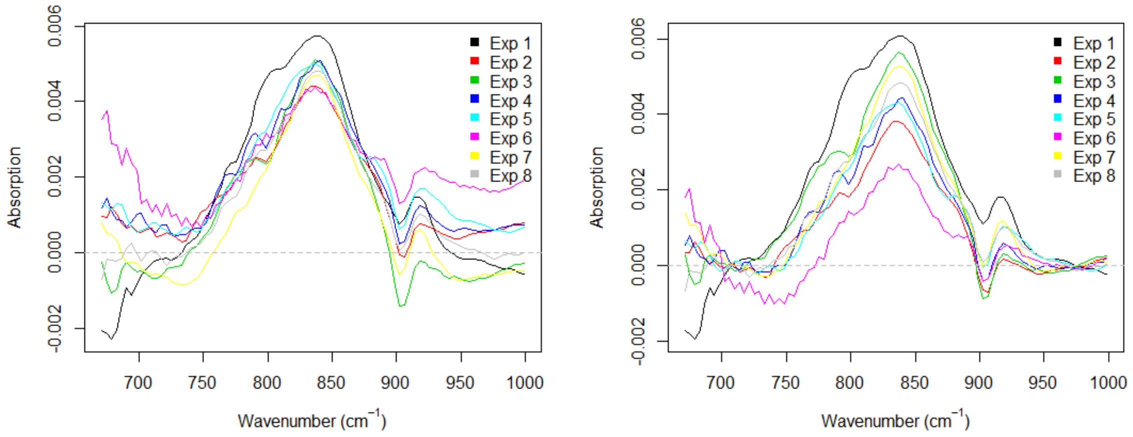
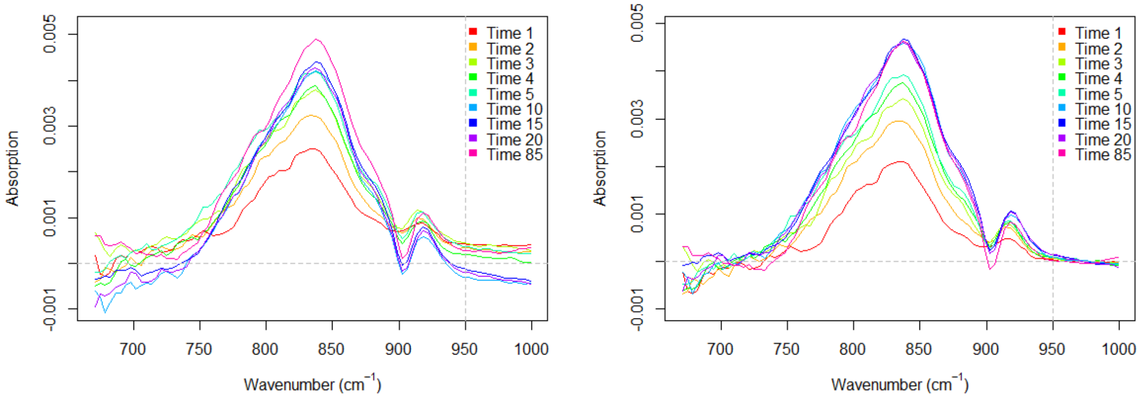


Figure 7.2: Baseline Adjustments Averaged at each Time



7.2 Theoretical Data Processing

The theoretical data provides calculated wavenumbers at which a particular chemical bond will vibrate, μ, along with its corresponding intensity, γ. In general, a simple quantum mechanical treatment of bond vibrations known as the harmonic oscillator model assumes the chemical bond to be a spring with a certain force constant that oscillates at a given frequency, and does not break regardless of how much the spring is stretched [21]. Realistically, the bond breaks if it is stretched far enough from its equilibrium distance. The anharmonic model accounts for this and predicts the resulting change in bond energy [21, 22, 23]. This deviation from harmonicity is accounted for by scaling the theoretical frequencies by some common percentage. We have verified the acceptable scaling range to be between $0\% - 10\%$. The scaling of these theoretical values, μ, for a given model cluster results in better comparisons with experimental frequencies using the real molecule.

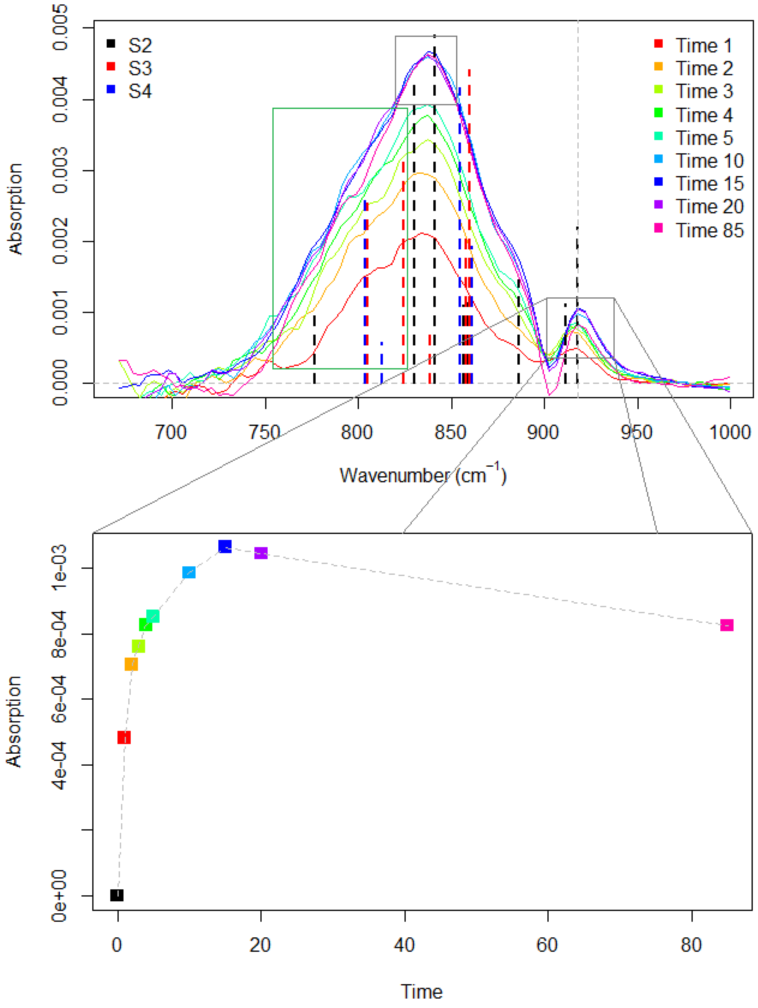
From a number of experimental studies using ATR-FTIR [33, 34, 37], evidence suggests that S_2 has the greatest presence throughout the entire experiment (namely, from $t = 0min$ to $t = 85min$) but that S_4 will very gradually dominate the mixture in the long run. In particular, we have verified that at $t = 85min$, the concentrations will hold the relationship $X_{2,85} > X_{3,85} > X_{4,85}$ (as discussed in section 5.2).

To reflect this, we scale the 17 theoretical wavenumbers by first focusing on S_2 in isolation. By examining the right most peak in Figure 7.3, we see the trend in its modes overtime exhibit a sharp increase followed by a gradual decrease. This trend is strongly consistent with the reaction system (1) in that that S_2 gains majority presence at earlier times but naturally starts to gradually diminish as stronger bonds begin to form at later times.

By also aligning its high intensity wavenumbers with the central peak, this would allow S_2 to claim the greatest absorptions and thus further reflect the belief that S_2 is dominant.

For illustration, we scale all theoretical wavenumbers by 5.5% (×1.055) in order to align the locations of S_2 theoretical wavenumbers with 1) the right most experimental peak and 2) the central peak.

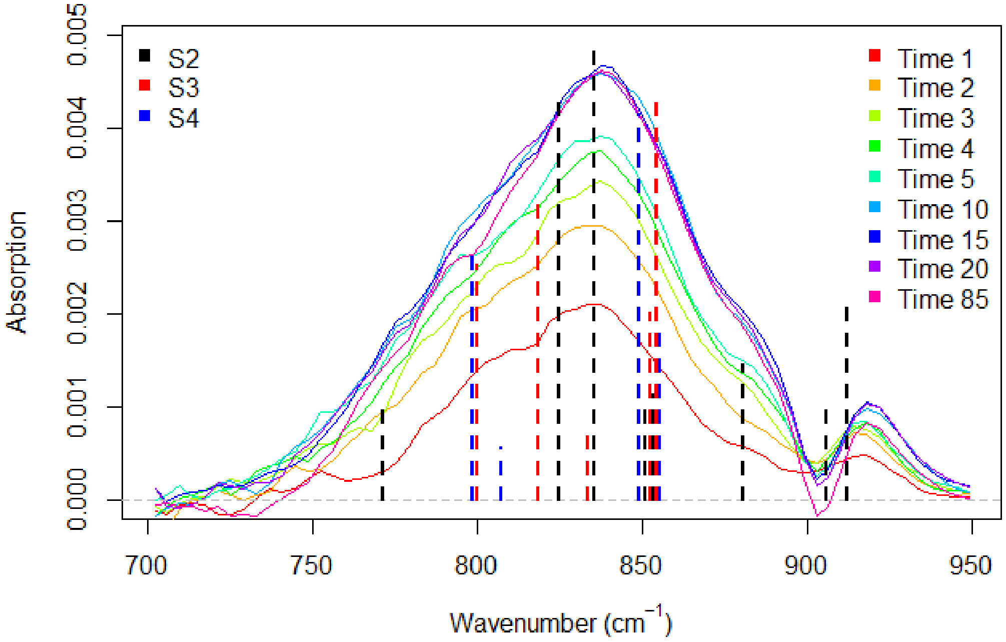
Figure 7.3: S_2 Scaling Analysis



However, also noting the green boxed area in Figure 7.3, we see that scaling the wavenumbers by a factor of 5.5% leaves very few theoretical wavenumbers accounting for the entire left side of the main peak which may indicate that the scaling is too high. We finally choose a scaling factor of 4.8% to have more wavenumbers account for the left side of the main peak and still have the two S_2 wavenumbers close enough to the rightmost peak.

Lastly, we eliminate all experimental wavenumbers $\omega < 700$ and $\omega > 950$. As these are understood to be non-absorbing ranges, we have verified that any fluctuations or patterns in those ranges are purely machine noise. The final data includes 65 absorption readings for wavenumbers $700 \leq \omega \leq 950$ and theoretical wavenumbers scaled at 4.8% ($\times 1.048$).

Figure 7.4: Final Processed Data



7.3 Estimation

To estimate the model parameters Θ we might consider sampling from the joint posterior density $p(\Theta|\mathcal{D})$ by sequentially sampling each conditional posterior density

$$\begin{aligned}
&\Theta_1 \sim p(\Theta_1|\mathcal{D})\\
&\Theta_2 \sim p(\Theta_2|\Theta_1, \mathcal{D})\\
&\vdots\\
&\Theta_d \sim p(\Theta_d|\Theta_{d-1}, \Theta_{d-2}, \ldots, \Theta_1, \mathcal{D})
\end{aligned}$$

using a suitable MCMC algorithm such as the Gibbs sampler. However close examination of the log joint posterior density in Section 4.4 suggests that the above conditional posterior distributions would be very difficult to derive analytically. Moreover, even when reducing the parameter space to estimate a much simpler model, the Gibbs sampler resulted in several hours of runtime under simulated data. We instead sample from $p(\Theta|\mathcal{D})$ using Stan software; a probabilistic programming language for Bayesian inference which uses an efficient implementation of Hamiltonian Monte Carlo (HMC) in C++ [24, 25, 26].

Here we provide estimates for both candidate reaction systems suggested by (1)-(2); the 3 reaction system and the 4 reaction system. As seen through the simulation study, the model was very sensitive to σ inputs which would lead the sampling algorithm exploring regions of implausibly high σ values and incorrect ODE parameters. Moreover, the model was significantly more sensitive with respect to σ when estimating under real data even under strongly informative prior densities. As such, we focus on sampling $(\kappa, X_{1,0})$ under fixed plausible values $(\hat{\sigma}, \hat{\beta}, \hat{\tau})$.

As mentioned in Section 7.1, the data is processed to correct for misalignment along the absorption axis. As such, the processed data assumes a baseline of 0 across all experiments at all time points. Further, we non-parametrically estimate a noise variance parameter from the experimental data. We set

$$\hat{\beta} = (0, 0, 0, 0, 0, 0, 0, 0, 0)$$
$$\hat{\tau} = 0.0002$$

Using these values, estimating the parameters of interest is done in the following two steps.

i. **Estimate σ.** This step estimates plausible σ values that are likely to correspond to plausible $(\kappa, X_{1,0})$. This is done by choosing $(\kappa, X_{1,0}) = (\hat{\kappa}, \hat{X}_{1,0})$ which generate concentrations that are consistent with a priori beliefs as discussed in Section 5. Given such ODE parameters, σ is estimated as the posterior mean after sampling

$$\sigma \sim p(\sigma | \hat{\kappa}, \hat{X}_0, \hat{\beta}, \hat{\tau}, \mathcal{D})$$
$$\pi(\sigma) \sim \mathcal{W}eibull(\delta = 2, \xi = 10)$$

ii. **Estimate $(\kappa, X_{1,0})$.** Given $\hat{\sigma}$ from the previous step, the ODE parameters are estimated as the posterior mean after sampling

$$(\kappa, X_{1,0}) \sim p(\kappa, X_{1,0} | \hat{\sigma}, \hat{\beta}, \hat{\tau}, \mathcal{D})$$
$$\pi(\kappa, X_{1,0}) \sim g(\kappa)h(\kappa)\pi(X_{1,0})$$

7.3.1 4 Reaction Model Fit

Sampling from $p(\sigma | \hat{\kappa}, \hat{X}_{1,0}, \hat{\beta}, \hat{\tau}, \mathcal{D})$ in Step 1, we obtain the following posterior mean estimates $\hat{\sigma} = (24, 22, 61, 64, 11, 9, 90, 18, 23, 58, 11, 7, 12, 15, 16, 16, 17)$. In Step 2, we sample from $p(\kappa, X_{1,0} | \hat{\sigma}, \hat{\beta}, \hat{\tau}, \mathcal{D})$ with the prior specification that $[\Sigma_\kappa]_{jj} = [\Sigma_{\alpha^*}]_{jj} = \sigma_{X_{1,0}} = 1,000,000$; $\pi(\kappa, X_{1,0}) \propto 1$.

We obtain the following estimates of $(\kappa, X_{1,0})$ shown in Figure 7.5 with corresponding concentration process in Figure 7.6 and fitted absorption curves in Figure 7.7.

Figure 7.5: Estimated ODE Parameters

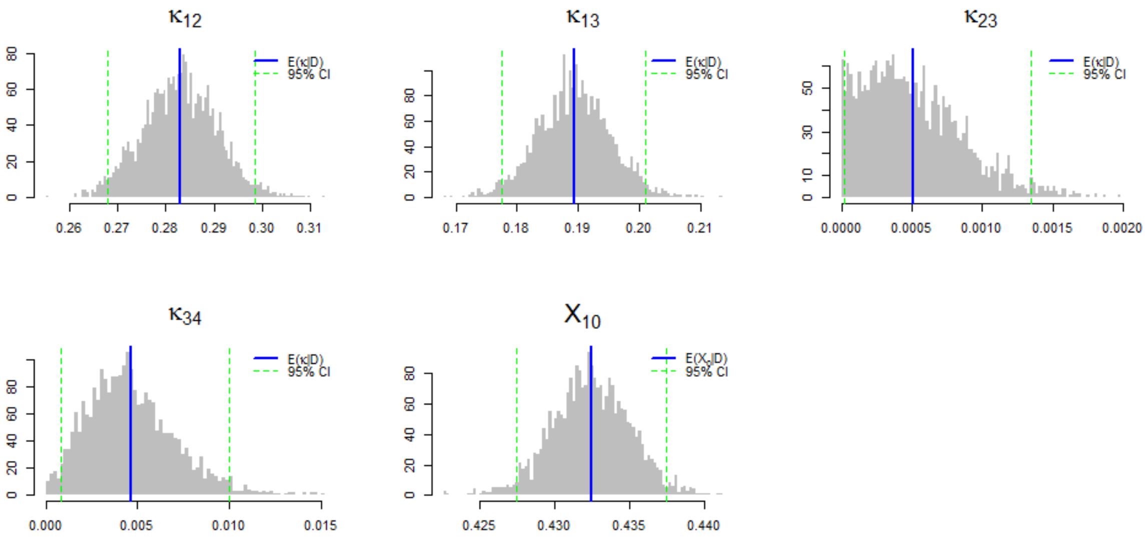


Figure 7.6: Estimated Concentration Curves

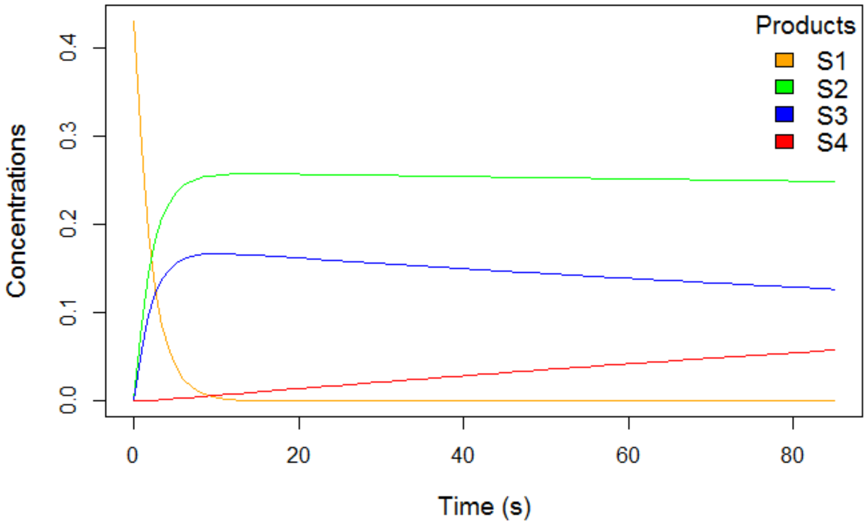
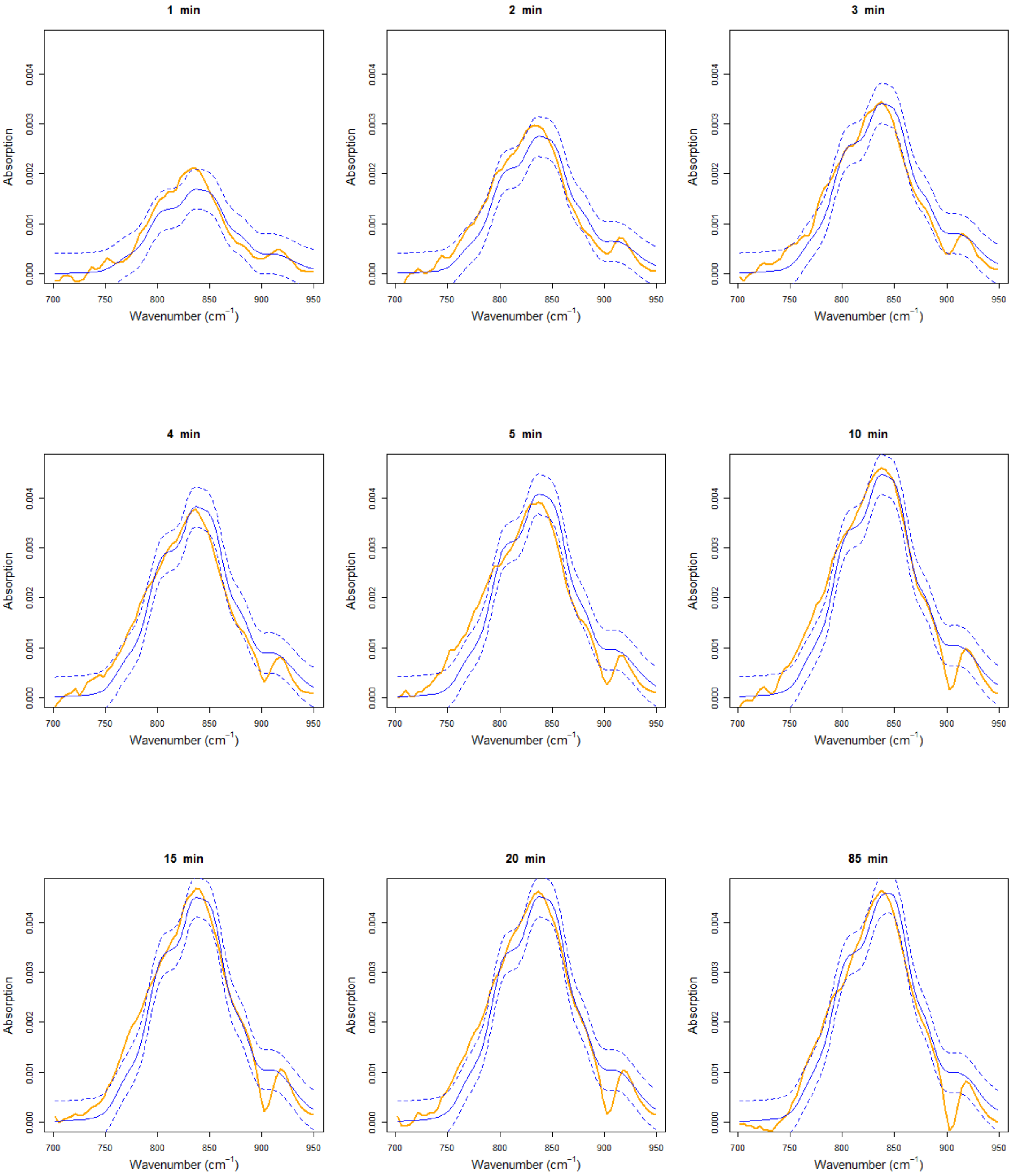


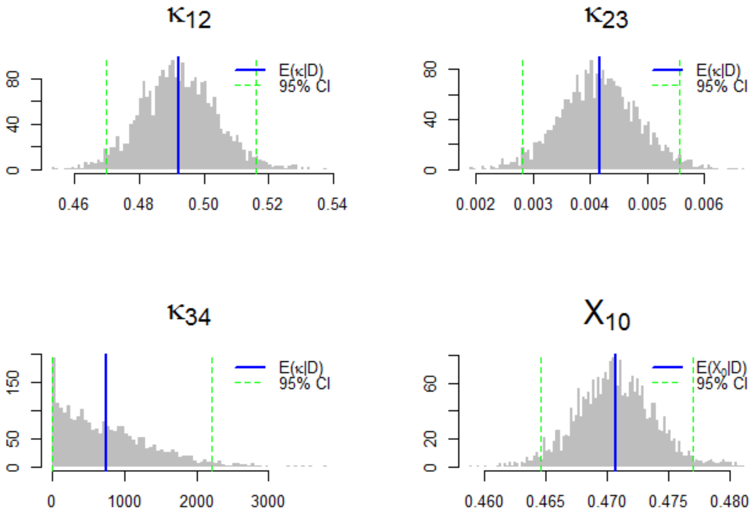
Figure 7.7: Estimated Absorption Curves



7.3.2 3 Reaction Model Fit

Sampling from $p(\sigma|\hat{\kappa}, \hat{X}_{1,0}, \hat{\beta}, \hat{\tau}, \mathcal{D})$ in Step 1, we obtain the following posterior mean estimates $\hat{\sigma} = (27, 31, 31, 50, 15, 78, 78, 97, 25, 17, 15, 7, 21, 20, 21, 16, 25)$. However when sampling from $p(\kappa, X_{1,0}|\hat{\sigma}, \hat{\beta}, \hat{\tau}, \mathcal{D})$ in Step 2, we see that the data gives very little information about κ_{34}. In particular, Figure 7.8 shows the estimates obtained for $(\kappa, X_{1,0})$ under a flat prior $\pi(\kappa, X_{1,0}) \propto 1$.

Figure 7.8: Estimated ODE Parameters



To obtain more plausible estimates, we impose the prior information: $\alpha_{2,85}(70\%) > \alpha_{3,85}(20\%) > \alpha_{4,85}(10\%)$. In order to obtain plausible estimates for κ, it turns out we require heavy reliance on this prior information. In particular, we set $\pi(\kappa) = g(\kappa)$ (defined in 5.2) where

$$\mu_{\alpha^*} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix}, \Sigma_{\alpha^*} = \begin{bmatrix} 0.05^2 & 0 & 0 \\ 0 & 0.05^2 & 0 \\ 0 & 0 & 0.05^2 \end{bmatrix}$$

We obtain the following estimates of $(\kappa, X_{1,0})$ shown in Figure 7.9 with corresponding concentration process in Figure 7.10 and fitted absorption curves in Figure 7.11.

Figure 7.9: Estimated ODE Parameters

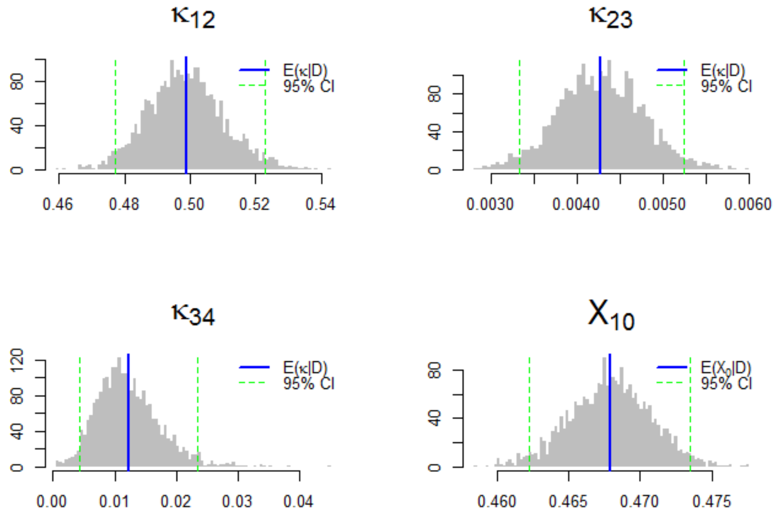


Figure 7.10: Estimated Concentration Curves

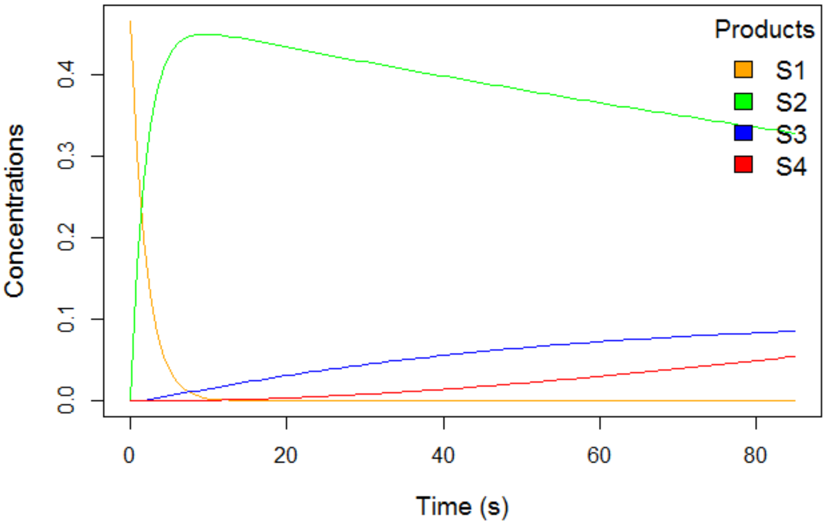
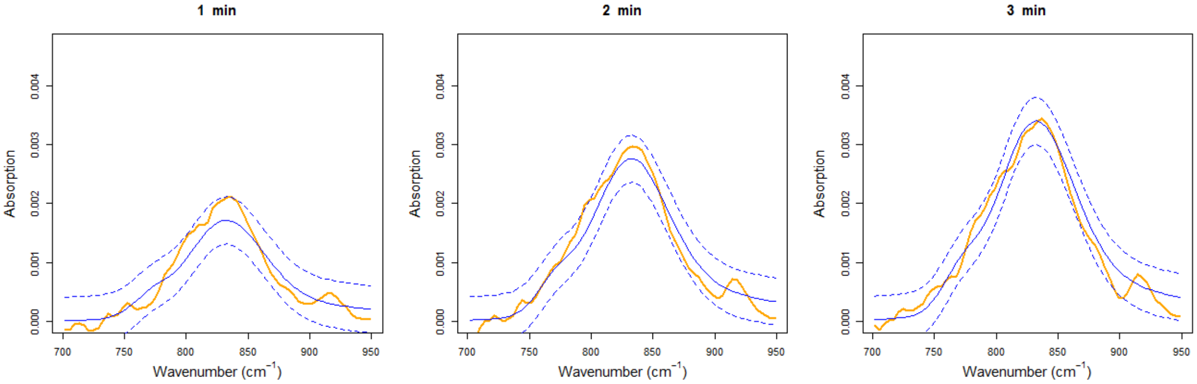
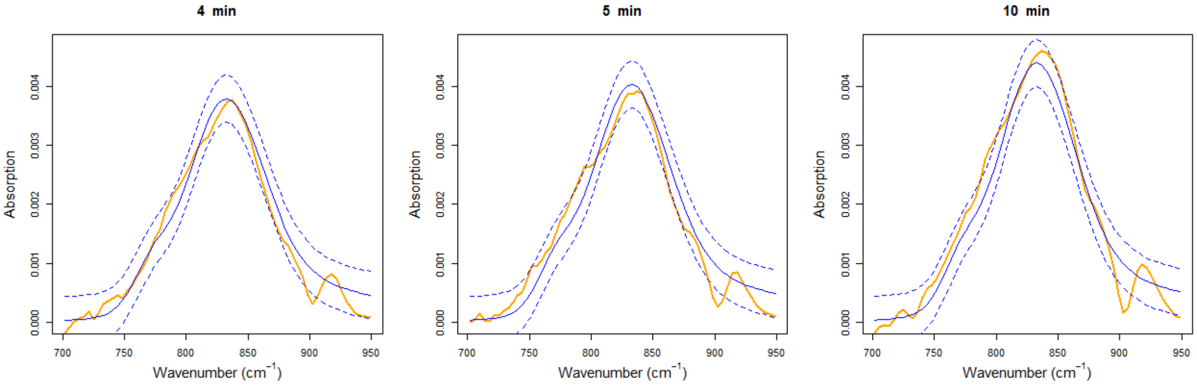
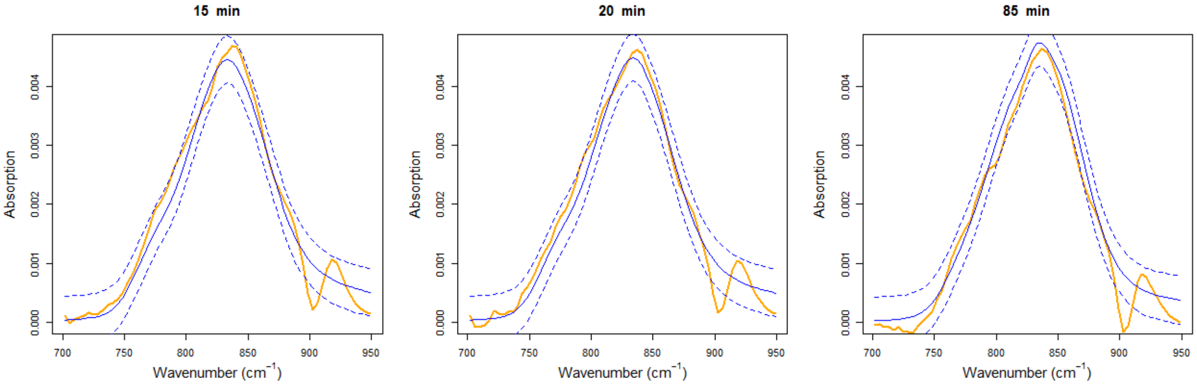


Figure 7.11: Estimated Absorption Curves




4 min
5 min
10 min
Absorption
0.000
0.001
0.002
0.003
0.004
700
750
800
850
900
950
Wavenumber (cm^{-1})


15 min
20 min
85 min
Absorption
0.000
0.001
0.002
0.003
0.004
700
750
800
850
900
950
Wavenumber (cm^{-1})

7.4 Model Comparison

Recall the measurement error model specified in Section 4.2:

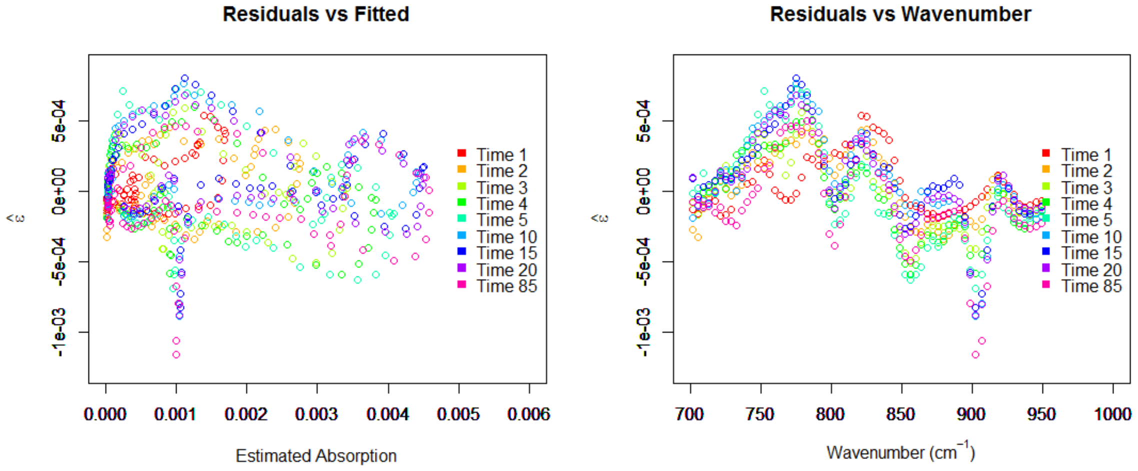
$$\epsilon_{it} \sim \mathcal{N}ormal(0, \tau^2) \qquad \forall i, \forall t$$

We analyze the residual diagnostic plots obtained from both model fits and provide a qualitative review of each model.

7.4.1 4 Reaction Model Diagnostics

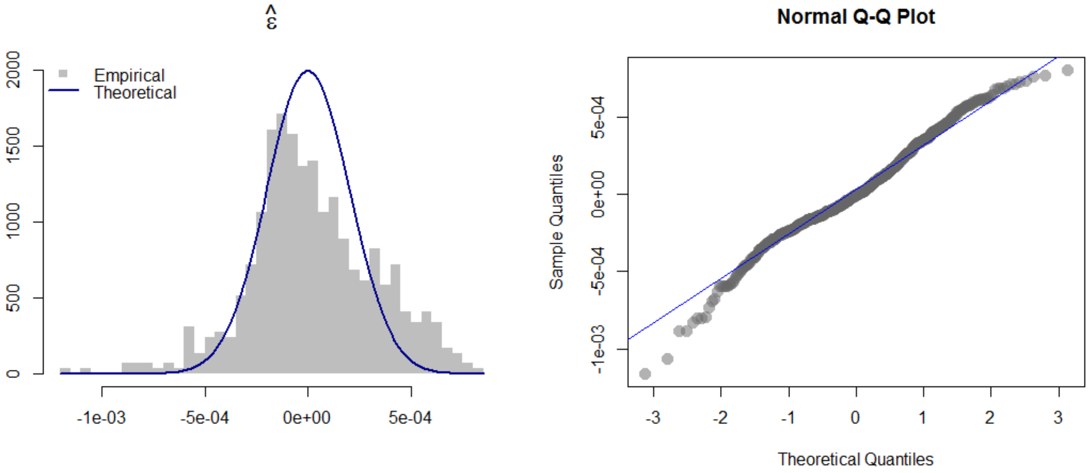
Figure 7.12 shows the Residuals plotted against both estimated absorption values and the wavenumber range. These plots suggest that the residuals are not completely random as specified by the measurement error model assumption. In particular, the residuals exhibit consistent oscillation patterns indicating that the model tends to fluctuate between underestimating and overestimating the absorption measures.

Figure 7.12: Model Residual Plots



As seen in Figure 7.13, the residuals clearly violate the normality assumption as both the histogram and Q-Q plot strongly suggest a rightly skewed residual distribution.

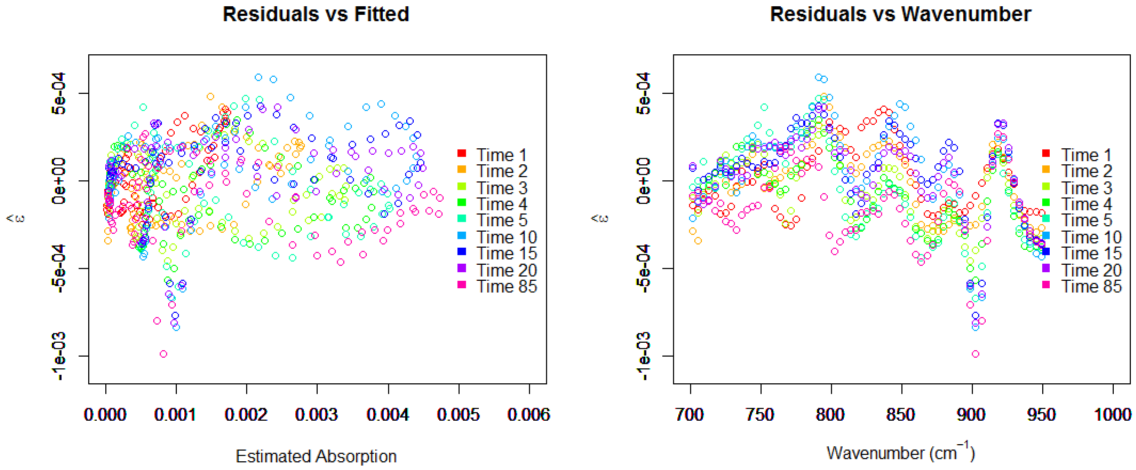
Figure 7.13: Model Residual Diagnostics



7.4.2 3 Reaction Model Diagnostics

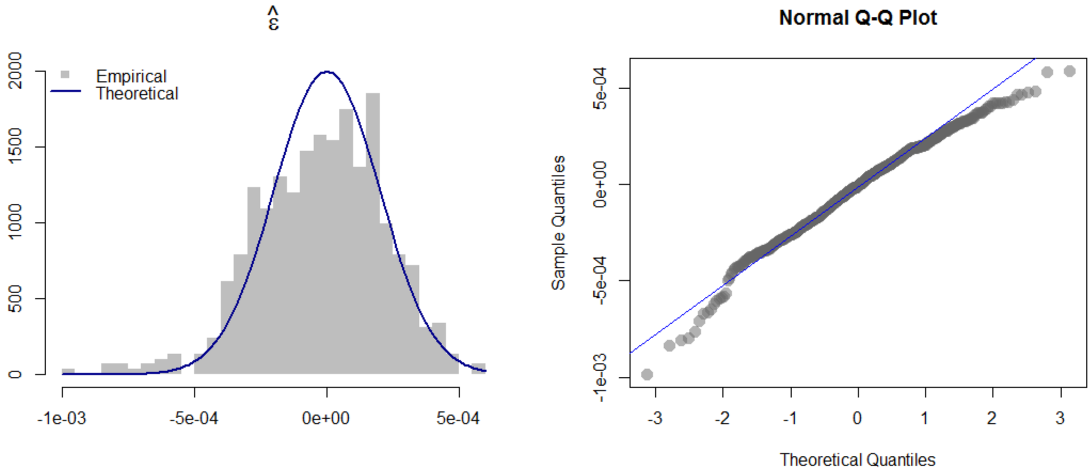
Figure 7.14 shows the Residuals plotted against both estimated absorption values and the wavenumber range. These plots also suggest that the residuals are not completely random as specified by the measurement error model assumption. As in the 4 reaction model, the residuals exhibit consistent oscillation patterns however appear to be slightly more scattered.

Figure 7.14: Model Residual Plots



As seen in Figure 7.15, the residuals clearly violate the normality assumption as both the histogram and Q-Q plot strongly suggest rightly skewed residuals. In comparison to the 4 reaction model, however, the residuals show noticeably less severity in the violation as both the center and tails of the empirical distribution show less deviation from normality.

Figure 7.15: Model Residual Diagnostics



7.4.3 Qualitative Comparison

As shown in the analysis of the residuals, both model residuals exhibit non randomness; in particular we see consistent oscillation patterns which suggest that both models may not be capturing certain peaks in the absorption curves. Whether certain peaks pertain to legitimate absorption characteristics or just noise is difficult to distinguish, however some detailed analysis of the experimental data enables us to find common peaks across different curves which we assume are genuine. We illustrate some of these peaks in Figure 7.16 and discuss the ability of each model to reflect these peaks.

Figure 7.16: Key Features Observed in Data

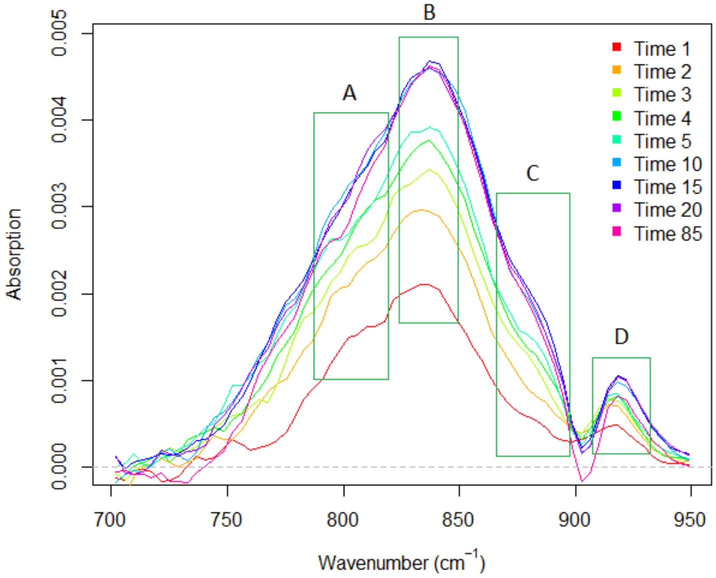
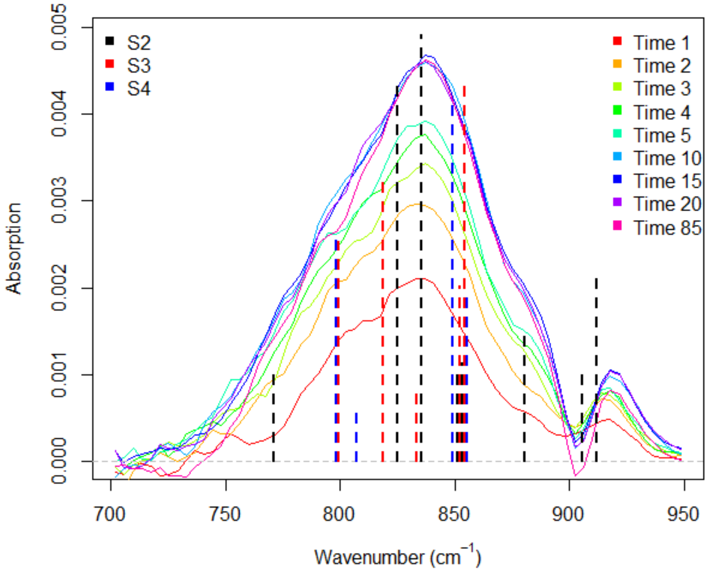


Figure 7.17: Final Processed Data



Region A emphasizes a slight bump in the absorption curve. This bump, though very subtle and slightly variable in location over time, appears on virtually all time points. Referring to Figure 7.17, we see that the scaling factor of 4.8% aligns this bump with the theoretical locations of S_3 and S_4. As seen in the estimated concentration curves for the 4 reaction model Figure 7.6, the model describes the chemical system as being significantly composed of S_3 at all times and thus is generally able to pick up these bumps at all time points. On the other hand, the concentration curves produced by the 3 reaction model Figure 7.10 suggest the model describes the system as having very little presence of both S_3 and S_4 at all times and is thus unable to pick up this bump even under strong prior specifications.

Region B is the main peak which is assumed to be attributed mainly to the dominance of S_2; both models adequately account for this main peak.

Region C is another subtle but consistent detail in the data; this bump exhibits a slight convexity and is attributed to S_2 given the scaling factor of 4.8%. As seen by the model fits for the 4 reaction model Figure 7.7, the model does in fact show an estimated convexity in that region, suggesting that the model adequately identifies that peak as legitimate rather than noise. The 3 reaction model fits Figure 7.11 however, does not identify this convexity at all regardless of the model describing the system as being mainly composed of S_2.

Region D is attributed to S_2 given the scaling factor; this peak shows an obvious presence at all time points suggesting that S_2 is present at all times throughout the experiment. The peak is reasonably accounted for by the 4 reaction model Figure 7.7, but not accounted for at all by the 3 reaction model Figure 7.11.

In addition to the comparisons outlined above, it is worth mentioning that the $\hat{\sigma}$ estimates obtained in the estimation process for the 3 reaction model are on average higher than the estimates corresponding to the 4 reaction model. This suggests that the 3 reaction model estimates flatter (hence less plausible) absorption peaks from the data to account for the lack of S_3 and S_4 estimated by the model which is undesirable. The 4 reaction model generally outputs lower $\hat{\sigma}$ estimates and attributes the aggregate absorption curve to a more plausible balance of the three species.

Lastly, both models are able to estimate plausible parameters from the data that are consistent with a priori beliefs. However, the 4 reaction model is able to estimate these parameters given no prior specification of these beliefs, $\pi(\kappa, X_{1,0}) \propto 1$, suggesting that this model is reflective of the true absorption process. On the other hand, the 3 reaction model requires heavy reliance on prior information which might suggest that it is incorrectly specified.

Chapter 8

Discussion

As there is generally no statistical model that gives forms to the spectra of each species by directly relating reaction rate constants to Infrared Spectroscopy data, the model proposed in this thesis provides a fresh statistical perspective to the problem of inferring rate constants. Although this model serves as a starting point for future work, it already adds a parametric interpretation of the IR Spectroscopy which the conventional model-free MCR methods do not provide; moreover it provides this under a drastically reduced parameter set. Further, this model can be formulated to reflect different reaction systems (first order and higher order reactions) assess evidence a posteriori for or against a set of candidate governing reaction systems to a particular chemical system of interest.

We propose the 4 reaction model is more suitable in describing the chemical mixture of interest. In addition the reasons outlined in Section 7.4.3, the Bayesian Inference provides us with credible intervals for κ which we interpret as evidence against the 3 reaction model being suitable. Referring to their corresponding reaction channels (1)-(2) we see that the 3 reaction system is simply the 4 reaction system reduced by the channel

$$S_1 \xrightarrow{\kappa_{13}} S_3$$

Figure 7.5 illustrates 95 % credible intervals which provide strong evidence a posteriori that $\kappa_{13} \neq 0$. As such, we believe this reaction is statistically significant in describing the underlying chemical process.

Future work may consider several extensions. The first extension would be to estimate a reasonable σ simultaneously with the ODE parameters, however, this estimation is currently limited by the data provided. As an illustration, revisit Figure 7.4 and carefully examine the spacing between the experimental wavenumbers and the spacing between the ordered theoretical wavenumbers. It can be seen that in some cases, $(|\omega_{i+1} - \omega_i| > |\mu_{(j+1)} - \mu_{(j)}|)$ making certain Gaussian components virtually indistinguishable from others.

A second extension may be to also account for the individual experimental trials, l, which would significantly increase the amount of evidence for parameter estimates a posteriori at the cost of only adding a few more parameters under the current model structure (intercepts and initial concentrations).

$$\mathcal{A}_t(\omega_i) = \beta_{lt} + \sum_{j=2}^{4} f_j(t; \kappa, X_0^l) \sum_{k=1}^{|S_j|} \gamma_{jk} \phi(\omega_i; \mu_{jk}, \sigma_{jk})$$

$$\forall i, \forall l, \forall t$$

Thus reflecting the number of experiments in the statistical model of measurement error

$$\epsilon_{ilt} \overset{iid}{\sim} \mathcal{N}ormal\left(0, \tau^2\right)$$

$$\forall i, \forall l, \forall t$$

So far, all inference has been done after averaging all experimental data at each time to obtain smoother absorption curves. However, the efficacy of this extension would depend on the quality of the experimental data; Figure 3.4 shows that certain experiments are far too variable (with regards to noise and vertical shift) and thus may lead to very volatile model outputs. Alternatively, the current model may show less sensitivity given smoother averaged curves, however obtaining smoother average curves would require many more experimental runs which may be very costly for scientists.

A third extension may consider modeling the absorption data as a Poisson count of photon absorption, given data of photon absorption counts

$$\tilde{\mathcal{A}}_{ilt} \sim \mathcal{P}oisson(\mathcal{A}_{ilt}) \approx \tilde{\mathcal{A}}_{ilt} \sim \mathcal{N}ormal(\mathcal{A}_{ilt}, \mathcal{A}_{ilt})$$

References

[1] Julia Tofan-Lazar, Hind A. Al-Abadleh. *ATR-FTIR Studies on the Adsorption/Desorption Kinetics of Dimethylarsinic Acid on Iron-(Oxyhydr)oxides.* The Journal of Physical Chemistry, 2012.

[2] William Mitchell, Sabine Goldberg, Hind A. Al-Abadleh. *In situ ATR-FTIR and surface complexation modeling studies on the adsorption of dimethylarsinic acid and p-arsanilic acid on iron-(oxyhydr)oxides.* Journal of Colloid and Interface Science, 2011.

[3] Jim Clark `http://www.chemguide.co.uk/analysis/uvvisible/beerlambert.html`. 2007.

[4] Materials Evaluation and Engineering Inc `http://www.mee-inc.com/hamm/fourier-transform-infrared-spectroscopy-ftir/`. Fourier Transform Infrared spectroscopy (FTIR).

[5] `http://chemwiki.ucdavis.edu/Physical_Chemistry/Spectroscopy/Vibrational_Spectroscopy/Infrared_Spectroscopy/Infrared%3A_Interpretation`. UC Davis

[6] C. Ruckebusch, L. Blanchet *Multivariate curve resolution: A review of advanced and tailored applications and challenges.* Analytica Chimica Acta, 2013.

[7] Multivariate Curve Resolution Homepage `http://www.mcrals.info/`. Webpage of the MCR-ALS method with programs, tutorials and datasets

[8] J. Saurina, S. Hernandez-Cassau, R. Tauler, A. Izquierdo-Ridorsa.
Multivariate Resolution of Rank-Deficient Spectrophotometric Data from First-Order Kinetic Decomposition Reactions. Journal of Chemometrics, 1998.

[9] MCR-ALS-Theory
`http://www.cid.csic.es/homes/rtaqam/tmp/WEB_MCR/mcrals.html`. Centre d'Investigacio i Desenvolupament; Consell Superior d'Investigacions Cientifiques

[10] Roma Tauler `http://www.cid.csic.es/homes/rtaqam/tmp/WEB_MCR/download/pdf/MCR_2005.pdf`. Centre d'Investigacio i Desenvolupament; Consell Superior d'Investigacions Cientifiques, 2005.

[11] Henning Schroder, Mathias Sawall, Christoph Kubis, Detlef Selent, Dieter Hess, Robert Franke, Armin Borner, Klaus Neymeyr
On the ambiguity of the reaction rate constants in multivari ate curve resolution for first-order reaction systems. Submitted 2015.

[12] J LS Lee, I S Gilmore: A Guide to the Practical Use of Multivariate Analysis in SIMS `http://www.simssociety.org/PPT/IanGilmore/Gilmore2_MVATutorial_a.pdf` National Physical Laboratory, Teddington, UK

[13] Luca Cardelli.
From Processes to ODEs by Chemistry. Microsoft Research.

[14] H Finotti.
Math 231:Introduction to Ordinary Differential Equations- Mini-Project: Modeling Chemical Reaction Mechanisms Department of Mathematics, University of Tennessee. Fall 2012.

[15] Marcel Maeder, Yorck-Michael Neuhold.
Practical Data Analysis in Chemistry. Chapter 3, Pages 82-83.

[16] `http://chemwiki.ucdavis.edu/Physical_Chemistry/Kinetics/Reaction_Rates/First-Order_Reactions` UC Davis.

[17] http://chemwiki.ucdavis.edu/Physical_Chemistry/Spectroscopy/Vibrational_Spectroscopy/Infrared_Spectroscopy UC Davis.

[18] C.-P. Sherman Hsu.
Infrared Spectroscopy. Separation Sciences Research and Product Development. Mallinckrodt, Inc. Mallinckrodt Baker Division

[19] Praveen Kumar Mogili
Heterogeneous Chemistry and Extinction Measurements of Mineral Dust Components PhD Thesis, University of Iowa, 2007.

[20] Algorithms used for Microspectroscopy.
https://www.microspectra.com/support/service-contracts/algorithms-used-for-microspectroscopy. CRAIC Technologies Mobile.

[21] Brian Smith.
Infrared Spectral Interpretation: A Systematic Approach. CRC Press, 1998. P. 15-20.

[22] Mark E. Tuckerman.
Bond vibrations. Advanced Chemisttry, Lecture 18. New York University.

[23] Anharmonic Oscillator. http://chemwiki.ucdavis.edu/Physical_Chemistry/Quantum_Mechanics/06._One_Dimensional_Harmonic_Oscillator/Harmonic_Oscillator/Anharmonic_Oscillator. UC Davis.

[24] stan-software:2015.
Stan: A C++ Library for Probability and Sampling, Version 2.9.0. http://mc-stan.org/. 2015.

[25] stan-manual:2015
Stan Modeling Language Users Guide and Reference Manual, Version 2.9.0. http://mc-stan.org/ 2015.

[26] rstan-software:2015

RStan: the R interface to Stan, Version 2.8.0 `http://mc-stan.org/rstan.html` 2015.

[27] Brian E. Blank, Steven George Krantz.

Calculus: Single Variable, Volume 1. KeyCollege Publishing. Springer. 2006.

Section 3.3 P 177.

[28] Math 20 - Introduction to Linear Algebra and Multivariable Calculus

Chapter 5 Eigenvalues and Eigenvectors. `http://www.math.harvard.edu/archive/20_spring_05/handouts/ch05_notes.pdf` Harvard Mathematics Department. Spring 2005.

[29] Robert A. Beezer.

A First Course in Linear Algebra. Properties of Eigenvalues and Eigenvectors. `http://linear.ups.edu/html/section-PEE.html` 2015.

[30] Diagonalizable matrices. `http://s-mat-pcs.oulu.fi/∼mpa/matreng/ematr4_2.htm` Mathematics Division, University of Oulu, Finland.

[31] Andrew Gelman, Frederic Bois, Jiming Jiang.

Physiological Pharmacokinetic Analysis using Population Modeling and Informative Prior Distributions. Journal of the American Statistical Association. 1996.

[32] J. O. Ramsay, G. Hooker, D. Campbell and J. Cao.

Parameter Estimation for Differential Equations: A Generalized Smoothing Approach J. R. Statist. Soc. B 69, Part 5, pp. 741–796. 2007.

[33] Tofan-Lazar, J.; Al-Abadleh, H.A., Atr-ftir

studies on the adsorption/desorption kinetics of dimethylarsinic acid on iron-(oxyhydr)oxides. J. Phys. Chem. A 2012, 116, 1596-1604.

[34] Tofan-Lazar, J.; Al-Abadleh, H.A.,

Kinetic atr-ftir studies on phosphate adsorption on iron-(oxyhydr)oxides in the absence *and presence of surface arsenic: Molecular-level insights into the ligand exchange mechanism.* J. Phys. Chem. A 2012, 116, 10143-10149.

[35] Adamescu, A.; Hamilton, I.P.; Al-Abadleh, H.A.,
Thermodynamics of dimethylarsinic acid and arsenate interactions with hydrated iron-(oxyhydr)oxide clusters: DFT calculations. Environ. Sci. Technol. 2011, 45, 10438-10444.

[36] Adamescu, A.; Mitchell, W.; Hamilton, I.P.; Al-Abadleh, H.A.,
Insights into the surface complexation of dimethylarsinic acid on iron (oxyhydr)oxides from ATR-FTIR studies and quantum chemical calculations. Environ. Sci. Technol. 2010, 44, 7802-7807.

[37] Sabur, M.A.; Goldberg, S.; Gale, A.; Kabengi, N.J.; Al-Abadleh, H.A.,
Temperature-dependent ATR-FTIR and calorimetric studies on arsenicals adsorption from solution to hematite nanoparticles. Langmuir 2015, 31, 2749-2760.

[38] Anna de Juan, Marcel Maeder, Manuel Martinez, Roma Tauler
Combining hard- and soft-modelling to solve kinetic problems Chemometrics and Intelligent Laboratory Systems 54 2000 123–141

[39] Arsenic `http://www.who.int/mediacentre/factsheets/fs372/en/` World Health Organization

Other References

[40] Sabine Bijlsma
Estimating rate constants of chemical reactions using spectroscopy. PhD Thesis. University of Amsterdam, Amsterdam, The Netherlands, 2000.

[41] Wentzell, Peter D., et al.
Multivariate curve resolution of time course microarray data. BMC bioinformatics 7.1 (2006): 343.

[42] Esteban, M., et al.

Multivariate curve resolution with alternating least squares optimisation: a soft-modelling approach to metal complexation studies by voltammetric techniques. TrAC Trends in Analytical Chemistry 19.1 (2000): 49-61.

[43] Jaumot, Joaquim, Anna de Juan, and Roma Tauler.

MCR-ALS GUI 2.0: New features and applications. Chemometrics and Intelligent Laboratory Systems 140 (2015): 1-12.

Appendices

A. Aggregate Reaction

Theorem: Suppose $f(x) = \sum_i f_i(x)$. Provided that $f_i'(x)$ exists $\forall i$, then by linearity of differentiation [27]

$$\frac{d}{dx} f(x) = \frac{d}{dx} \sum_i f_i(x) = \sum_i \frac{d}{dx} f_i(x)$$

Recall the 4 reaction concentration ODE system (1)

$$\frac{d}{dt} X_{1t} = -(\kappa_{12} + \kappa_{13}) X_{1t} \tag{8.1}$$

$$\frac{d}{dt} X_{2t} = \kappa_{12} X_{1t} - \kappa_{23} X_{2t} \tag{8.2}$$

$$\frac{d}{dt} X_{3t} = \kappa_{13} X_{1t} + \kappa_{23} X_{2t} - \kappa_{34} X_{3t} \tag{8.3}$$

$$\frac{d}{dt} X_{4t} = \kappa_{34} X_{3t} \tag{8.4}$$

and consider the aggregate concentration $X_{At} = X_{2t} + X_{3t} + X_{4t}$. By Theorem above, we have

$$\begin{aligned} \frac{d}{dt} X_{At} &= \frac{d}{dt}\Big(X_{2t} + X_{3t} + X_{4t}\Big) \\ &= \frac{d}{dt} X_{2t} + \frac{d}{dt} X_{3t} + \frac{d}{dt} X_{4t} \end{aligned}$$

Thus

$$\begin{aligned}\frac{d}{dt}X_{At} &= (12) + (13) + (14)\\ &= \kappa_{12}X_{1t} - \kappa_{23}X_{2t}\\ &+ \kappa_{13}X_{1t} + \kappa_{23}X_{2t} - \kappa_{34}X_{3t}\\ &+ \kappa_{34}X_{3t}\\ &= (\kappa_{12} + \kappa_{13})X_{1t}\end{aligned}$$

This leads to the reduced pairwise ODE system

$$\begin{aligned}\frac{d}{dt}X_{1t} &= -(\kappa_{12} + \kappa_{13})X_{1t}\\ \frac{d}{dt}X_{At} &= (\kappa_{12} + \kappa_{13})X_{1t}\end{aligned}$$

which implies only one reaction channel

$$S_1 \xrightarrow{\kappa_{12}+\kappa_{13}} S_A$$

B. Eigenvalues and Eigenvectors

Theorem 1: A scalar λ is an eigenvalue of an $n \times n$ matrix Ω if and only if λ satisfies the characteristic equation [28]

$$\det(\Omega - \lambda I) = 0$$

With respect to the **4 reaction system**, it can be seen by inspection that its corresponding ODE system (3) can be written in matrix form $X_t' = \Omega X_t$. In particular we have

$$X_t' = \begin{bmatrix} \frac{d}{dt}X_{1t} \\ \frac{d}{dt}X_{2t} \\ \frac{d}{dt}X_{3t} \\ \frac{d}{dt}X_{4t} \end{bmatrix} = \begin{bmatrix} -(\kappa_{12}+\kappa_{13}) & 0 & 0 & 0 \\ \kappa_{12} & -\kappa_{23} & 0 & 0 \\ \kappa_{13} & \kappa_{23} & -\kappa_{34} & 0 \\ 0 & 0 & \kappa_{34} & 0 \end{bmatrix} \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \\ X_{4t} \end{bmatrix} = \Omega X_t$$

Note that due to Ω being lower triangular, $\det(\Omega - \lambda I)$ is simply the product of the diagonal entries of $\Omega - \lambda I$, thus obtaining the characteristic polynomial

$$(\lambda + \kappa_{12} + \kappa_{13})(\lambda + \kappa_{23})(\lambda + \kappa_{34})\lambda = 0$$

which implies the following unique real valued eigenvalues

$$\begin{aligned} \lambda_1 &= -(\kappa_{12} + \kappa_{13}) \\ \lambda_2 &= -\kappa_{23} \\ \lambda_3 &= -\kappa_{34} \\ \lambda_4 &= 0 \end{aligned}$$

Further solving $(\Omega - \lambda_q I)\vec{v} = 0 \ \forall q = 1, 2, 3, 4$ separately we obtain the following real valued eigenvectors:

$$\vec{v}_1|\lambda_1 = \begin{pmatrix} \frac{(\kappa_{12}+\kappa_{13}-\kappa_{23})(\kappa_{12}+\kappa_{13}-\kappa_{34})}{(\kappa_{13}-\kappa_{23})\kappa_{34}} \\ -\frac{(\kappa_{12}+\kappa_{13}-\kappa_{34})\kappa_{12}}{(\kappa_{13}-\kappa_{23})\kappa_{34}} \\ -\frac{\kappa_{12}+\kappa_{13}}{\kappa_{34}} \\ 1 \end{pmatrix}, \vec{v}_2|\lambda_2 = \begin{pmatrix} 0 \\ -\frac{\kappa_{34}-\kappa_{23}}{\kappa_{34}} \\ -\frac{\kappa_{23}}{\kappa_{34}} \\ 1 \end{pmatrix}, \vec{v}_3|\lambda_3 = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \vec{v}_4|\lambda_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Similarly for the **3 reaction system**, we can see that its ODE system (4) can be expressed in matrix form as

$$\begin{bmatrix} \frac{d}{dt}X_{1t} \\ \frac{d}{dt}X_{2t} \\ \frac{d}{dt}X_{3t} \\ \frac{d}{dt}X_{4t} \end{bmatrix} = \begin{bmatrix} -\kappa_{12} & 0 & 0 & 0 \\ \kappa_{12} & -\kappa_{23} & 0 & 0 \\ 0 & \kappa_{23} & -\kappa_{34} & 0 \\ 0 & 0 & \kappa_{34} & 0 \end{bmatrix} \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \\ X_{4t} \end{bmatrix}$$

By Theorem 1 and applying similar Eigen decomposition as for the 4 reaction system above, we have the following eigenvalues and eigenvectors for the 3 reaction system:

$$\lambda_1 = -\kappa_{12}$$

$$\lambda_2 = -\kappa_{23}$$

$$\lambda_3 = -\kappa_{34}$$

$$\lambda_4 = 0$$

$$\vec{v}_1|\lambda_1 = \begin{pmatrix} -\frac{(\kappa_{12}-\kappa_{23})(\kappa_{12}-\kappa_{34})}{\kappa_{23}\kappa_{34}} \\ \frac{\kappa_{12}(\kappa_{12}-\kappa_{34})}{\kappa_{23}\kappa_{34}} \\ -\frac{\kappa_{12}}{\kappa_{34}} \\ 1 \end{pmatrix}, \vec{v}_2|\lambda_2 = \begin{pmatrix} 0 \\ -\frac{\kappa_{34}-\kappa_{23}}{\kappa_{34}} \\ -\frac{\kappa_{23}}{\kappa_{34}} \\ 1 \end{pmatrix}, \vec{v}_3|\lambda_3 = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \vec{v}_4|\lambda_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

C. (i) ODE Solution

Theorem 2: Suppose that Ω is an $n \times n$ square matrix and $S = \{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n\}$ is a set of eigenvectors with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$. If $\lambda_i \neq \lambda_j$, $i \neq j$ then S is a linearly independent set [29].

Theorem 3: An $n \times n$ matrix Ω is diagonalizable if and only if Ω has n linearly independent eigenvectors [30].

Noting that $\lambda_1 \neq \lambda_2 \neq \lambda_3 \neq \lambda_4$ in Appendix B, then by the above Theorem 2, $\vec{v}_1, \vec{v}_2, \vec{v}_3, \vec{v}_4$ is a linearly independent set. Further by Theorem 3, we know that Ω is diagonalizable such that it can be decomposed into the product of three $n \times n$ matrices

$$\Omega = Q \Lambda Q^{-1}$$

where $Q = [\vec{v}_1, \vec{v}_2, \vec{v}_3, \vec{v}_4] \in \mathbb{R}^{4\times 4}$ is a matrix of eigenvectors of Ω and $\Lambda = diag(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \in \mathbb{R}^{4\times 4}$.

We now find the general solution, X_t, to the ODE system

$$X_t' = \Omega X_t$$

By inspection, we see that the solution to this differential equation is the exponential function

$$X_t = e^{\Omega t} \vec{c}$$

where $\vec{c} \in \mathbb{R}^4$ is some vector independent of t. Rewriting the solution as the infinite Taylor series expansion of the exponential function about $t = 0$ (namely, the Maclaurin series) and applying the diagonalization theorem we obtain

$$\begin{aligned}X_t &= e^{\Omega t}\vec{c}\\ &= \left[\sum_{i=0}^{\infty}\frac{(\Omega t)^i}{i!}\right]\vec{c}\\ &= \left[\sum_{i=0}^{\infty}\frac{(Q\Lambda Q^{-1})^i t^i}{i!}\right]\vec{c}\end{aligned}$$

Note that since $Q^{-1}Q = I$, $\forall i \in \mathbb{N}$ we have

$$\begin{aligned}(Q\Lambda Q^{-1})^i &= Q\Lambda Q^{-1} \times Q\Lambda Q^{-1} \times Q \ldots Q^{-1} \times Q\Lambda Q^{-1}\\ &= Q\Lambda I\Lambda I \ldots I\Lambda Q^{-1}\\ &= Q\Lambda^i Q^{-1}\end{aligned}$$

Thus,

$$\begin{aligned}\left[\sum_{i=0}^{\infty}\frac{(Q\Lambda Q^{-1})^i t^i}{i!}\right]\vec{c} &= \left[\sum_{i=0}^{\infty}Q\frac{\Lambda^i t^i}{i!}Q^{-1}\right]\vec{c}\\ &= Q\left[\sum_{i=0}^{\infty}\frac{\Lambda^i t^i}{i!}\right]Q^{-1}\vec{c}\\ &= Q\left[\sum_{i=0}^{\infty}\frac{(\Lambda t)^i}{i!}\right]Q^{-1}\vec{c}\\ &= Qe^{\Lambda t}Q^{-1}\vec{c}\end{aligned}$$

Therefore, $X_t = Qe^{\Lambda t}Q^{-1}\vec{c}$. To find $\vec{c}$, we substitute the initial condition X_0 when $t = 0$

$$X_0 = Qe^{\Lambda(0)}Q^{-1}\vec{c} = QIQ^{-1}\vec{c} = \vec{c}$$

Thus the solution $X_t \in \mathbb{R}^4$ to the system $X_t' = \Omega X_t$ is

$$X_t = Qe^{\Lambda t}Q^{-1}X_0$$

Note that $e^{\Lambda t} \in \mathbb{R}^{4\times 4}$ where $\left[e^{\Lambda t}\right]_{qq} = e^{\lambda_q t}$ and $\left[e^{\Lambda t}\right]_{qp} = 0\ \forall q \neq p$. To see this, again consider the Maclaurin series expansion of the exponential function

$$\begin{aligned}
e^{\Lambda t} &= \sum_{i=0}^{\infty} \frac{(\Lambda t)^i}{i!} \\
&= \sum_{i=0}^{\infty} \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}^i \frac{t^i}{i!} \\
&= \sum_{i=0}^{\infty} \begin{bmatrix} \lambda_1^i & 0 & 0 & 0 \\ 0 & \lambda_2^i & 0 & 0 \\ 0 & 0 & \lambda_3^i & 0 \\ 0 & 0 & 0 & \lambda_4^i \end{bmatrix} \frac{t^i}{i!} \\
&= \begin{bmatrix} \sum_{i=0}^{\infty} \frac{(\lambda_1 t)^i}{i!} & 0 & 0 & 0 \\ 0 & \sum_{i=0}^{\infty} \frac{(\lambda_2 t)^i}{i!} & 0 & 0 \\ 0 & 0 & \sum_{i=0}^{\infty} \frac{(\lambda_3 t)^i}{i!} & 0 \\ 0 & 0 & 0 & \sum_{i=0}^{\infty} \frac{(\lambda_4 t)^i}{i!} \end{bmatrix} \\
&= \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 & 0 \\ 0 & e^{\lambda_2 t} & 0 & 0 \\ 0 & 0 & e^{\lambda_3 t} & 0 \\ 0 & 0 & 0 & e^{\lambda_4 t} \end{bmatrix}
\end{aligned}$$

C. (ii) ODE Solution Proportionality

Here we explain why $X_t \neq \alpha_t$ in general, but rather $X_t \propto \alpha_t$. Recall the solution to the system $X_t' = \Omega X_t$ obtained in Appendix C (i), namely

$$X_t = Q e^{\Lambda t} Q^{-1} X_0$$

Suppose that we scale X_t by some scalar $c \in \mathbb{R}$ such that $\tilde{X}_t = cX_t$. Mathematically we can express this as

$$\tilde{X}_t = c \cdot X_t = c \cdot Q e^{\Lambda t} Q^{-1} X_0 = Q e^{\Lambda t} Q^{-1} \tilde{X}_0$$

where $\tilde{X}_0 = c \cdot X_0$.

It is easy to see that regardless of how X_t is scaled, the only parameter that scales accordingly is X_0 but κ **remains unchanged.** This is a very important result of the linear ODEs because when we infer κ from the IR Absorption curves, we do not require their underlying areas to correspond to the true concentrations $\{X_t\}$ in the experimental mixture because the scaled areas, $\{\alpha_t\} = \{cX_t\}$, will theoretically correspond to the same κ values. Thus by Beer-Lambert law, if $X_t \propto \alpha_t$ then by our model

$$\begin{aligned} \{X_t\} &\implies (X_0, \kappa) \\ \{\alpha_t\} &\implies (cX_0, \kappa) \end{aligned}$$

Key idea: IR Absorption contributions of each species and concentrations of each species both correspond to the same κ value. As such, since α_0 is not the main parameter of interest and κ is independent of initial conditions, we simply refer to α_0 as X_0 in order to avoid confusion with the notations introduced at the beginning of the paper.

D. κ scaling factor

Recall Section 3.1 which showed the reduction of reaction channels (1) to the single reaction channel (6) of the aggregated concentration system X_{At}. This reduction showed that the aggregated concentration process, X_{At}, grows at rate constant $\kappa_A = \kappa_{12} + \kappa_{13}$. This would suggest that the sum of reaction rate constants (7)-(8) are proportional to some estimate of the true aggregate rate up to some scalar; namely, $\hat{\kappa}_A = \zeta(\hat{\kappa}_{12} + \hat{\kappa}_{13})$.

Given the aggregate concentration process, $\{Y_{At}\}$ as shown in Figure 3.1, we estimate the aggregate reaction rate constant to be $\hat{\kappa}_A \approx 0.6$ as shown in Figure 8.1 with corresponding estimated concentration curve illustrated in Figure 8.2.

Figure 8.1: Posterior samples of κ_A

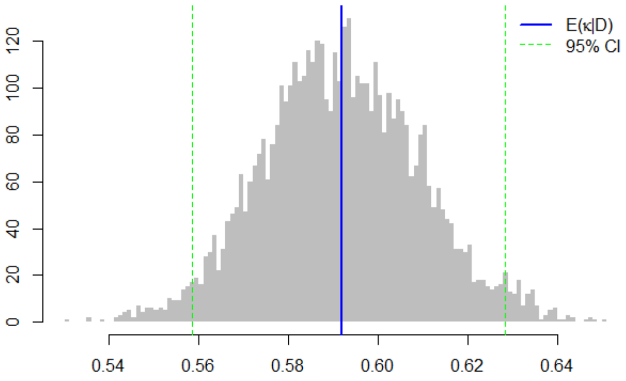
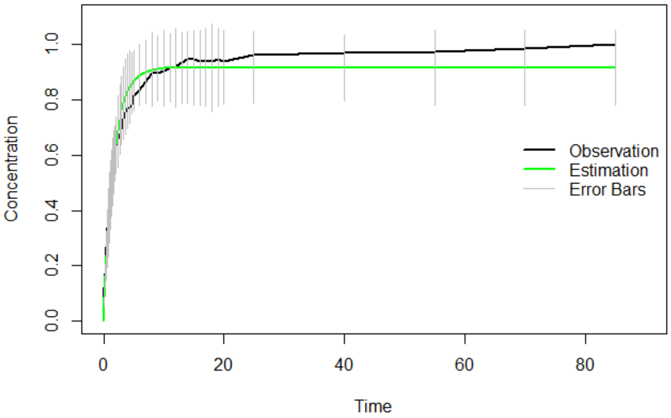


Figure 8.2: Estimated Aggregate Concentration Curve



$\hat{\kappa}_A \approx 0.6 \implies \zeta = \frac{0.6}{0.05+0.01} = 10$. Thus, scaling the given reaction rate constants (7)-(10) by $\zeta = 10$ yields the following adjusted a priori estimates:

$$\kappa_{12} : 0.5 \pm 0.1$$

$$\kappa_{13} : 0.1 \pm 0.1$$

$$\kappa_{23} : 0.1 \pm 0.1$$

$$\kappa_{34} : 0.01 \pm 0.005$$

E. Estimating $X_{1,0}$

Recall the chemical system reaction channels (1) dictate that at $t = 0$: $X_{1,0} > 0$ and $X_{2,0} = X_{3,0} = X_{4,0} = 0$. As such, defining the aggregate concentration curve as $X_{At} = X_{2t}+X_{3t}+X_{4t}$ we obtain initial conditions $(X_{1,0}, X_{A,0}) = (X_{1,0}, 0)$. Further, we obtain the following reaction channel (by Appendix A)

$$S_1 \xrightarrow{\eta} S_A$$

with corresponding ODE system

$$\begin{bmatrix} \frac{d}{dt}X_{1t} \\ \frac{d}{dt}X_{At} \end{bmatrix} = \begin{bmatrix} -\eta & 0 \\ \eta & 0 \end{bmatrix} \begin{bmatrix} X_{1t} \\ X_{At} \end{bmatrix}$$

By Appendix C (i), the solution to this system is of the form

$$\begin{aligned} X_t &= Qe^{\Lambda t}Q^{-1}X_0 \\ &= \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{-\eta t} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_{1,0} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} X_{1,0}e^{-\eta t} \\ X_{1,0}(1 - e^{-\eta t}) \end{bmatrix} = \begin{bmatrix} X_{1t} \\ X_{At} \end{bmatrix} \end{aligned}$$

$$\implies \lim_{t\to\infty} X_{At} = \lim_{t\to\infty} X_{1,0}(1 - e^{-\eta t}) = X_{1,0}$$

F. Statistical Model of Measurement Error

Recall from 3.2.1 that the observed absorption is defined as $\tilde{\mathcal{A}}_{it} = \log \frac{\tilde{I}^R_{it}}{\tilde{I}^S_{it}}$ (we drop the subscripts). Two points should be noted here: 1) Since the reference cell is assumed to be a fixed chemical surface, we can interpret $\tilde{I}_R$ as a constant photon intensity which stays fixed at each $\{i, t\}$ for all experiments. 2) The chemical surface is an iron-oxide which does not have significant absorption characteristics at the experimental wavenumbers.

As such, since $\tilde{I}_S$ is the stochastic component of $\tilde{\mathcal{A}}$ which can vary by experiment, and has significant absorption characteristics at the experimental wavenumbers, we define the stochastic photon intensity variable as $\tilde{I} = \frac{\tilde{I}_R}{\tilde{I}_S}$ which is considered very large as $\tilde{I}_R \gg \tilde{I}_S$.

Further, since the infrared spectrometer measures the amounts of photons absorbed at a given wavenumber, we can interpret the counts as a histogram where each bin represents the count of photons in each wavenumber bin. In particular, we assume a distribution over the counts at each $\{i, t\}$ as

$$\tilde{I} \sim \mathcal{P}oisson(I)$$

which can we approximated by the Normal distribution for very large I as

$$\tilde{I} \sim \mathcal{N}ormal(I, I)$$

However it is more reasonable to assume a smaller variance at each wavenumber, thus obtaining the scaled variance

$$\tilde{I} \sim \mathcal{N}ormal(I, \gamma^2 I)$$

If we define $\tilde{I} = IY$, it is easy to see that

$$Y \sim \mathcal{N}ormal\left(1, \frac{\gamma^2}{I}\right)$$

Therefore,

$$\begin{aligned}\log \tilde{I} &= \log IY \\ &= \log I + \log Y \\ &= \mathcal{A} + \log Y\end{aligned}$$

Define $g(Y) = \log Y$, by the Taylor Series expansion of $g(Y)$ about Y=1, we can obtain an approximation of the first moment

$$\begin{aligned}\mathbb{E}\Big[g(Y)\Big] &\approx g(\mu_Y) + \frac{g''(\mu_Y)}{2}\sigma_Y^2 \\ &= \log(\mu_Y) - \frac{1}{2\mu_Y^2}\sigma_Y^2 \\ &= \log(1) - \frac{1}{2 \cdot 1^2}\frac{\gamma^2}{I} \\ &\approx 0 && I \gg \gamma^2\end{aligned}$$

Further, the variance can be approximated by the delta method as

$$\begin{aligned}\mathbb{V}\Big[g(Y)\Big] &\approx \Big(g'(\mathbb{E}[Y])\Big)^2 \mathbb{V}[Y] \\ &= \left(\frac{1}{\mathbb{E}[Y]}\right)^2 \mathbb{V}[Y] \\ &= \left(\frac{1}{1}\right)^2 \frac{\gamma^2}{I} \\ &= \frac{\gamma^2}{I}\end{aligned}$$

Since we assume $I \gg \gamma^2$, we set $\mathbb{V}(\log Y) = \tau^2$ where τ^2 is some constant for all $\{i, t\}$. Thus,

$$\tilde{\mathcal{A}}_{it} = \mathcal{A}_{it} + \log Y_{it}$$

$$Y_{it} = \epsilon_{it} \sim \mathcal{N}ormal(0, \tau^2)$$

$$\forall i, \forall t$$