The Dissertation Committee for Keegan Hines
certifies that this is the approved version of the following dissertation:

Bayesian Approaches For Modeling Protein Biophysics

Committee:

Richard Aldrich, Supervisor

Lawrence Cormack

Daniel Johnston

Michael Mauk

Jonathan Pillow

Bayesian Approaches For Modeling Protein Biophysics

by

Keegan Hines, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2014

Dedicated to my wife. Thank you for everything.

Acknowledgments

I gratefully acknowledge the myriad people who helped me, in various ways, throughout graduate school. I would first like to thank my advisor, Rick Aldrich, for providing an ideal intellectual environment. Rick afforded me a great deal of scientific freedom to pursue and learn whatever I thought was interesting and important. Early on, this freedom was accompanied by inevitable failed projects, but in the end allowed me to develop a level of independence for which I will always be grateful. I have also learned to be fearless in unfamiliar areas and to constantly move outside of my comfort zone. Equally important, I am glad for the opportunities to break up work with, say, multi-hour discussions on the relative merits of early Pink Floyd albums or detailed lessons on the important developments in jazz in 1959.

I have been fortunate to have learned from, and worked closely with, each of the post-docs in the Aldrich lab. Tom Middendorf and I worked together for quite a while on a project that is described in Chapter 2. I will fondly remember Tom's contagious enthusiasm and joy in science and his unwavering optimism and determination. I worked with Brent Halling on a project involving Bayesian modeling of analytical ultracentrifugation data, which is not included here. Words such as *thorough* and *determined* can hardly begin to describe Brent, a tireless investigator who settles for nothing but his personal best. I have enjoyed working recently with Xixi Chen, as she has helped me with electrophysiology experiments. I have not only benefitted from her meticulous experimental expertise, but also her pragmatic and honest outlook. With Jenni Greeson-Bernier, I spent some time bringing Bayesian methodology to the analysis of some incredible experiments she's conducting. Jenni's determination in the face of obstacles is truly admirable and I am very grateful for many discussions of science and career. Additionally, I thank Norma Gonazales for her help over the years and Margaux Miller for her tenacious efforts for the benefit of the lab.

I would like to thank the members of my dissertation committee: Larry Cormack, Dan Johnston, Mike Mauk, and Jonathan Pillow. Not only have I received important guidance regarding both my research and general scientific interests, but I was also lucky to undertake coursework from each of you, which played a strong role in the development of my interests and expertise.

I am happy to have spent my time amongst good friends in the Institute for Neuroscience and will sorely miss Akram, Dean, Nick, Jeremy, Jake, Ben, Leor, Kenneth, Scott, Brooks, and many others. I am also grateful to the staff of the Institute for Neuroscience and the Center for Learning and Memory, people whose praises are not sung often enough, especially Krystal Phu, Wilbert King, and Chris Weatherly.

I'm happy to thank the American Heart Association for funding my graduate research.

I am also grateful to Jack Haar and Bill Guido, formerly of the Medical College of Virginia, for providing my first research experience in neurobiology and starting me on this path.

I thank my wife, Caitlin, for her endless support. Thank you for always being there to celebrate successes and accept failures. I happily dedicate this dissertation to you. And I thank my parents and siblings for being supportive of my endeavors throughout life, no less so now. We look forward to returning to the east coast and being closer to family.

Bayesian Approaches For Modeling Protein Biophysics

Publication No. ____________________

Keegan Hines, Ph.D.
The University of Texas at Austin, 2014

Supervisor: Richard Aldrich

Proteins are the fundamental unit of computation and signal processing in biological systems. A quantitative understanding of protein biophysics is of paramount importance, since even slight malfunction of proteins can lead to diverse and severe disease states. However, developing accurate and useful mechanistic models of protein function can be strikingly elusive. I demonstrate that the adoption of Bayesian statistical methods can greatly aid in modeling protein systems. I first discuss the pitfall of parameter non-identifiability and how a Bayesian approach to modeling can yield reliable and meaningful models of molecular systems. I then delve into a particular case of non-identifiability within the context of an emerging experimental technique called single molecule photobleaching. I show that the interpretation of this data is non-trivial and provide a rigorous inference model for the analysis of this pervasive experimental tool. Finally, I introduce the use of nonparametric Bayesian inference for the analysis of single molecule time series. These methods aim to circumvent problems of model selection and parameter identifiability and are demonstrated with diverse applications in single molecule biophysics. The adoption of sophisticated inference methods will lead to a more detailed understanding of biophysical systems.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Protein Biophysics

Proteins are the fundamental unit of computation and signal processing in biological systems. The coordinated actions of protein systems are responsible for complex cellular processes such as DNA replication/transcription, muscle contraction, and action potential generation. Even slight mutations in proteins can lead to severe disease states including epilepsy (Caterall, 2012), cardiac arrhythmia (Splawski et al., 2000), cystic fibrosis (Vankeerberghen et al., 1998), sickle cell disease (Higgs et al., 1989), and muscular dystrophy (Roberts et al., 1992), among very many others. Therefore, it is vitally important to develop a detailed and mechanistic understanding of protein function, and malfunction, in order to develop effective treatments and therapies for diverse diseases.

The quantitative modeling of biophysical systems has a rich history, with several exemplary investigations throughout the 20th century. In the 1950s, Alan Hodgkin and Andrew Huxley embarked on a series of studies to understand the ionic basis of the action potential (Hodgkin and Huxley, 1952). In this work, they argued for the existence of distinct voltage-gated conductances within axon membranes, which turn on and off in response to voltage fluctuations and which act in concert to shape the action potential. In their model, they imagined that each of these conductances might exist in one of two states (conductive and non-conductive) and the transition rate between these states is altered by transmembrane voltage. Through careful and thorough experimentation, they were able to measure the voltage dependence of each of these conductances and then combine these into a general dynamical system model which could accurately explain the evolution of voltages and conductances underlying the action potential.

In earlier work, Lenor Michaelis and Maud Menten sought to develop a quantitative theory of enzyme kinetics (Michaelis and Menten, 1913). In their model, an enzyme E and a substrate S evolved toward an enzymatic product P according to the kinetic scheme,

$$E + S \rightleftharpoons ES \xrightarrow{k_{cat}} E + P. \tag{1.1}$$

In the limit where substrate concentration is far in excess of enzyme concentration, the rate of product formation is given by,

$$\frac{dP}{dt} = k_{cat} E \frac{S}{K_m + S}, \tag{1.2}$$

where K_m is the substrate concentration at which $\frac{dP}{dt}$ is half maximal. Thus, the product formation rate increases asymptotically toward its maximum value of $k_{cat}E$. This model has been widely successful to explain the kinetic parameters K_m and k_{cat} in a variety of biochemical systems.

More recently, Jacques Monod, Jeffries Wyman and Jean-Pierre Changeaux sought to understand the molecular basis of cooperativity in ligand binding systems (Monod et al., 1965). In this work, they explain binding cooperativity through a model of concerted allostery (see Figure 1.1). In this model, we imagine that a molecule can access one of two conformational states, denoted T and R. Both states are accessible to the protein in the absence of ligand, but one state is energetically preferred by a factor of L. Regulation of the protein by the ligand is achieved by supposing that the ligand binds preferably to the T state than to the R state. That is, the ligand binding affinity for each state is denoted k_R and k_T and we suppose that $k_T > k_R$. Therefore, the presence of ligand shifts the equilibrium between states R and T toward the state with higher ligand affinity. The effect of ligand on the equilibrium distribution of states R and T is called allostery and, in the model, is equal to the ratio of affinities, k_R/k_T. Importantly, this model of ligand binding to a single protein can be easily extended to the case of a homomeric receptor of multiple subunits. Here, the state equilibrium constant L is changed by a factor of $(k_R/k_T)^2, (k_R/k_T)^3, (k_R/k_T)^4$ for a dimer, trimer, and tetramer receptor, and so on. This model of cyclic allostery has been used to model many systems including hemoglobin (Ackers, et al., 1992), ligand-gated ion channels (Karlin, 1967; Lape et al., 2008), and voltage-gated ion channels (Marks and Jones, 1992; Zagotta et al., 1994; Horrigan and Aldrich, 2002), among many others.

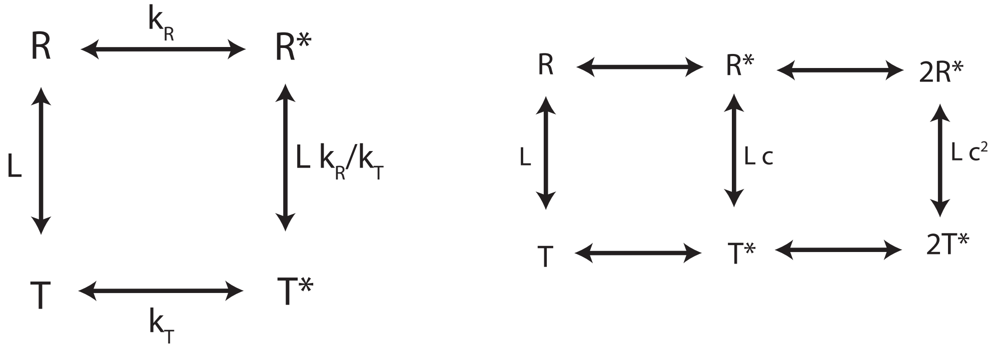

Figure 1.1: Cyclic allosteric model of Monod, Wyman, and Changeux. (Left) In this model, the protein can exist in one of two states, R or T, both of which it can access in the presence or absence of ligand. The equilibrium constant between the two states (in the absence of ligand) is denoted L. It is assumed that ligands bind each state with distinct affinity such that $k_T > k_R$. Therefore, the presence of ligand favors the T state by a factor proportional to $c = k_T/k_R$. (Right) This model can be extended to multisite receptors and provides a simple explanation of cooperative binding. Note that subsequent binding events alter the $R \leftrightarrow T$ transition by an additional factor of k_T/k_R.

These three examples, which sit amongst numerous others, demonstrate features that are common to many quantitative models of biomolecules. Very generally, we imagine that the system of interest has access to some fixed number of biophysically relevant states which are connected in some particular arrangement. The dynamics of the system between these states is governed by transition rates linking each state. Models of this kind belong to a general class called *state space models* (SSMs) which are widely used to describe many phenomena (Oppenheim and Schafer, 1999). With using an SSM to model a biomolecular system, we are generally aiming to learn three things about the system: the states, the connectivity between the states, and the transition rates between the states. When SSMs are interpreted within a biophysical context, the system states are thought to correspond to energetically semi-stable conformational states of the protein and the transition dynamics governed by free-energy barriers and the principles of statistical physics. In this setting, state transitions occur in a memoryless way, simply depending on free-energy barriers separating states. For this reason, SSMs of biomolecules can be thought of as discrete-state Markov processes. The task for the experimentalist is to use the tools of Markov theory, and whatever experimental manipulations are available, to develop accurate Markovian models and estimate relevant transition rates. As will be discussed throughout the following chapters, the task of using data to develop meaningful models can be non-trivial as systems of higher complexity are studied.

1.2 Statistical Inference

Statistics is quantified knowing. Figure 1.2 is a representation of the process of modeling systems in the world. At left, our system of interest undergoes some dynamics which are hidden from us; the system exists in a black box. However, we can make some measurements of the system. Inevitably though, we can only measure some, never all, of the properties (or degree of freedom) of the system. Therefore, any measurement induces a coarse-graining or otherwise obfuscation of the true system dynamics. Our task is then to use this imperfect information in order to construct the best possible model of the true underlying system. Given some measurement we have made about the world, it is the tools of statistical inference that allow us to rigorously quantify exactly what we do and do not know. It is only statistics that shines a light on the relationship between the things we would like know, and the things we can confidently assert.

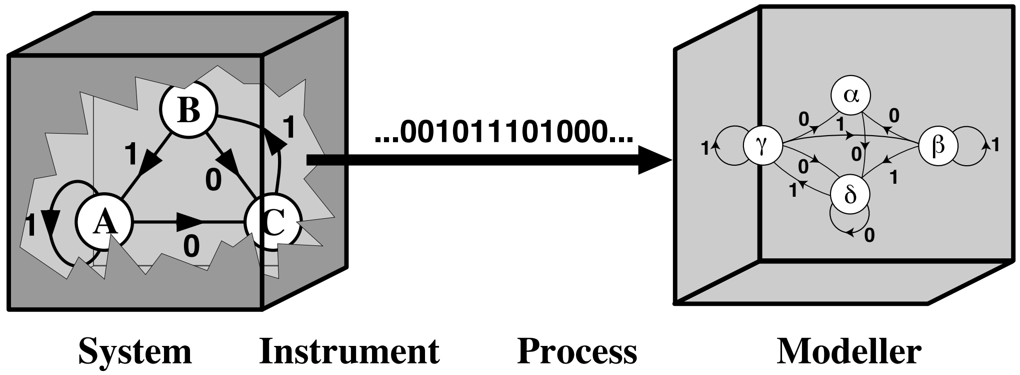

Figure 1.2: Schematic of the process of modeling natural systems. At left, we have our system of interest which undergoes some dynamics which are unavoidably hidden from us; the system exists in a black box. Making measurements of the system allows us access to some, but certainty not all, of the system's degrees of freedom and thus induces a coarse-graining of the true properties of the system. The task as a modeler is to is to use the coarse-grained measurement to devise an accurate model of the true dynamics. Imagery stolen from Jim Crutchfield lecture notes.

Since much of what is described in the following chapters champions a Bayesian approach to the quantitative modeling of proteins, what follows is a brief overview of the history and philosophy of Bayesian inference as contrasted with more common techniques. The Frequentist approach to statistics gained wide popularity in the 20th century after the remarkable work of Ronald Fisher, Jerzy Neyman and others. Frequentists proposed that one could quantify uncertainty and confidence by appealing to a notion of repeated experimentation (Fisher, 1922; Neyman, 1939). The data we have, a frequentist might argue, is but one random draw from the set of all possible data we might have seen. Therefore, we can appeal to the notion of an infinite number of repeated experiments and ask how rare our data is compared to all possible draws. This supposition is incredibly powerful if we can make certain assumptions about the kinds of distributions from which our data were drawn. Thanks to the Central Limit Theorem, we have strong assurance that many kinds of random variables should be approximately Normally distributed, allowing Frequentists to derive elegant expressions for what various data draws ought to look like. From these ideas emerge the mainstays of common statistical usage: hypothesis tests, ANOVA, multiple regression. The generality and simplicity of frequentist methodologies have lead to their widespread adoption in nearly all areas of science and engineering.

An alternative philosophy of statistics was proposed by Thomas Bayes and fully developed by Pierre-Simon Laplace. In what history has termed Bayesian inference, an opposite approach is taken to quantifying uncertainty.

The Bayesian prefers not to appeal to the Frequentists' infinite repetition of experiments, but instead accepts the data as fixed and treats the parameters of interest as random. That is not to say that the parameters of the world are genuinely believed to be random, but instead that we treat them as random variables (given the data) in order to quantify uncertainty.

Figure 1.3 summarizes the basic viewpoints of these two camps with respect to their philosophy of statistics. At the top we see the Frequentist view: the parameters we seek are some fixed property of the world and the process of gathering finite data induces randomness. In this view, we need to consider $p(x|\theta)$, the probability of seeing various possible datasets x, given a fixed value of θ. The bottom of Figure 1.3 shows the Bayesian view, which is opposite. Here, we consider that the data are fixed, and are instead interested in the set of all possible θ that could have generated the data and a quantification thereof, $p(\theta|x)$.

From a simple manipulation of the definition of joint and conditional probability, we can arrive at Bayes' rule,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}. \tag{1.3}$$

This equation provides a method to calculate the *posterior distribution*, $p(\theta|x)$, which quantifies the probability distribution over parameters θ, given data x. The components of Bayes rule are: the likelihood $p(x|\theta)$, the prior distribution $p(\theta)$, and the marginal evidence $p(x)$. In common practice,

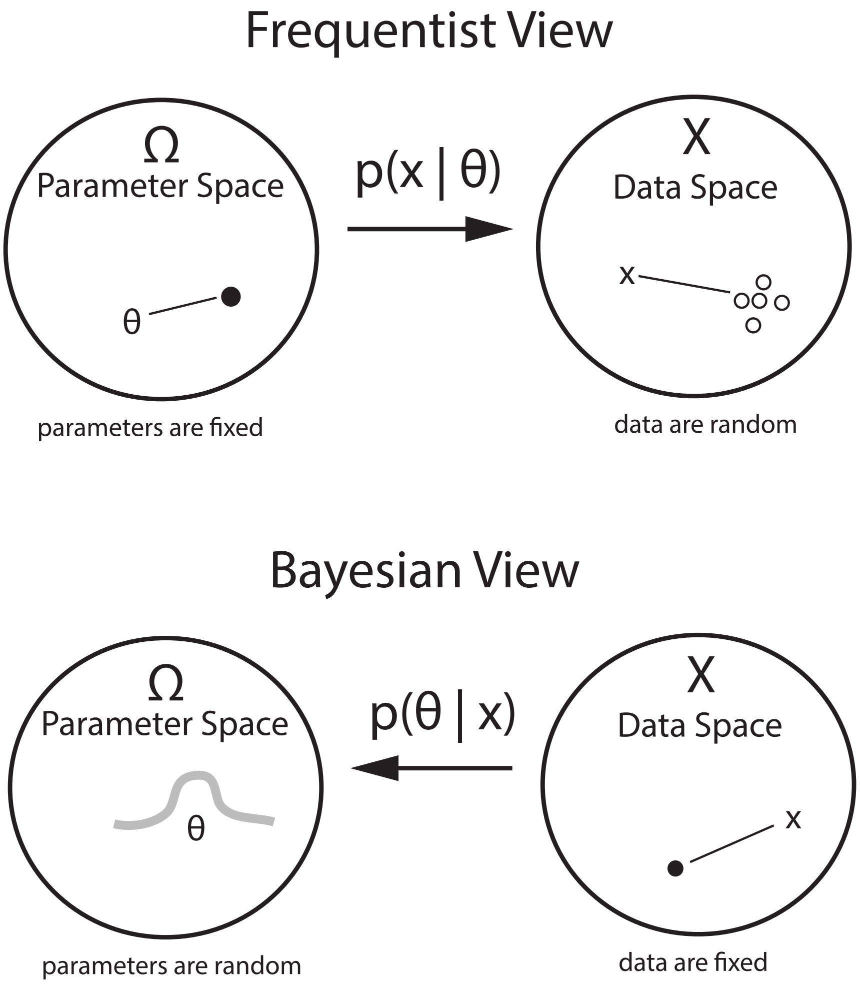

Figure 1.3: Diagram of Frequentist and Bayesian philosophy of statistics. Imagery adapted from Jonathan Pillow lecture notes.

we really only need to know the posterior distribution up to a constant of proportionality, so Bayes' rule is most commonly seen as,

$$p(\theta|x) \propto p(x|\theta)p(x). \tag{1.4}$$

Bayes rules states that the thing we're interested in, $p(\theta|x)$, is simply equal to the Frequentist's likelihood, $p(x|\theta)$, multiplied by a prior distribution, $p(\theta)$. The prior distribution quantifies our beliefs about θ before the experiment is performed. The prior distribution, which is inherently subjective, has been a source of contention between Frequentists and Bayesians, leading some Bayesians to expend effort to try to devise prior distributions which are as uninformative and objective as possible (Jeffrys, 1945). In fact, it was a rejection of the notion of incorporating prior beliefs into data analysis that left Bayesian methods in the background for the past few centuries. This battle became especially fierce with the demonstrated success of Frequentist methods in the early 20th century. I'll not belabor an extensive history of the ebb and flow of Bayesian methods, but a very entertaining account can be found in *The Theory That Would Not Die*, by Sharon B. McGrayne.

Despite continued resistance, Bayesian methods have had a resurgence in recent decades due to advances in computing power. Notice that Bayes' rule (equation 1.4) is hardly different than the Frequentist likelihood function, therefore it is hard to imagine situations where one method is clearly better or worse than the other. Recall that much of the utility of Frequentist methods is that under fairly general assumptions, we can derive analytical expressions for optimal estimates of parameters and confidence intervals. The Bayesian could do the same, incorporating prior beliefs in order to derive a simple form of the posterior distribution. However, when the models that are considered become arbitrarily complex, we won't necessarily be able to derive a simple form for Frequentist (or Bayesian) confidence intervals. The resolution for the Bayesian-minded investigator comes in the form of efficient algorithms for estimating posterior distributions. Known collectively as Markov chain Monte Carlo sampling, these methods rely on the strategy that we might approximate any arbitrarily complex probability distribution by drawing independent and identically distributed (iid) samples from it (Tierney, 1994). The posterior distribution, $p(\theta|x)$, will generally be very high-dimensional and without a simple closed form. Therefore, drawing iid samples from it might appear to be exactly as challenging as computing the posterior by brute force. However, MCMC achieves this goal by simulating a Markov chain whose limiting distribution is the posterior distribution of interest. Then, by simply simulating the Markov chain for some finite amount of time, we generate a finite number of iid samples from the target distribution.

Generating a Markov chain whose limiting distribution is some particular target distribution can be achieved in several ways. The first of these, now known as the Metropolis-Hastings algorithm, was proposed in the 1950s in order to approximate high-dimensional problems in particle physics (Metropolis et al., 1953). For simplicity, I'll describe a simple special case called the

Metropolis Random Walk. For iteration i of the algorithm, the Markov chain is in position θ_i in the parameter space. A potential transition to a new point, $\tilde{\theta}$, is generated by a random walk according to the following rules. The posterior probability of this potential point, $p(\tilde{\theta}|y_N)$, is computed and compared to the posterior probability of the current position of the chain, $p(\theta_i|y_N)$. If the proposal point has greater posterior probability than the current point, then it is accepted as a sample from the posterior distribution. If it has lower posterior probability, then it is rejected with probability proportional to the decrease in posterior probability. Thus, transitions of the Markov chain are accepted with probability α, where

$$\alpha = min\left(1, \frac{p(\tilde{\theta}|y_N)}{p(\theta_i|y_N)}\right). \tag{1.5}$$

This algorithm results in a Markov chain that explores the parameter space in proportion to the posterior probability. Therefore, the aggregate positions of the chain in each dimension provide iid samples from the corresponding marginal posterior distributions. This algorithm excels in its simplicity and its generality - we can use Metropolis-Hastings with any model for which we can compute posterior probability. In Chapter 2, I use this algorithm to perform Bayesian inference in a variety of biophysical settings including ligand binding models and dynamical systems.

An alternative implementation of MCMC, which I use in Chapter 4, is known as Gibbs sampling (Geman and Geman, 1984). Consider a general joint probability distribution between two random variables, $p(A, B)$. From the definition of conditional probability,

$$\begin{align} p(A|B) &= \frac{p(A, B)}{p(B)} \tag{1.6} \\ p(A, B) &= p(B)p(A|B) \tag{1.7} \\ p(A, B) &\propto p(A|B). \tag{1.8} \end{align}$$

Similarly, we could calculate the condition density with respect to the other variable,

$$\begin{align} p(B|A) &= \frac{p(A, B)}{p(A)} \tag{1.9} \\ p(A, B) &= p(A)p(B|A) \tag{1.10} \\ p(A, B) &\propto p(B|A). \tag{1.11} \end{align}$$

Thus, the joint distribution, $p(A, B)$, is linearly proportional to both conditional distributions, $p(A|B)$ and $p(B|A)$. This fact holds generally for joint distributions over any number of random variables and is the basis of Gibbs sampling. The strategy is that while the joint distribution, $p(A, B)$ might have no simple closed form, we can likely derive a simple form of each univariate conditional distribution. Generally, let $p(\theta_1, ..., \theta_K|x)$ be a K-dimensional posterior distribution with no simple closed form. If each univariate conditional distribution has a closed form such as $p(\theta_1|\theta_2, ..., \theta_K, x) \propto$

$F(\theta_1)$, then Gibbs sampling proceeds as following. For each iteration i of the algorithm, we draw the i^{th} random sample of each parameter according the univariate conditional distributions,

$$\theta_1^i \sim p(\theta_1|\theta_2^{i-1}, ..., \theta_K^{i-1}, x) = F(\theta_1) \tag{1.12}$$

$$\theta_2^i \sim p(\theta_2|\theta_1^{i}, ..., \theta_K^{i-1}, x) = F(\theta_2) \tag{1.13}$$

$$... \tag{1.14}$$

$$\theta_K^i \sim p(\theta_K|\theta_1^{i}, ..., \theta_{K-1}^{i-1}, x) = F(\theta_K). \tag{1.15}$$

Note that in this algorithm there is no accept/reject criterion since each sample is drawn iid from the corresponding conditional posterior distribution. This allows for better algorithm efficiency as compared to Metropolis-Hastings. Finally, I briefly mention other methods of MCMC, though they are not used here. Hamiltonian Monte Carlo (Neal, 2011) makes an analogy to the Hamiltonian energy function of classical physics in order to construct a Markov chain which explores posterior distributions faster and more thoroughly than Metropolis-Hastings and this approach has been generalized (Girolami and Calderhead, 2011). These sampling methods, and others, aim to efficiently explore complex posteriors by adapting to the local structure of the distributions. For high-dimensional biophysical models, I suspect that these adaptive methods will be required going forward.

It is important to appreciate what these computational methods make possible. In biophysics, detailed mechanistic models are evaluated based on their ability to explain carefully controlled experimentation. Given some data x and some model $p(x|\theta)$, we would like to use the data to gain a good estimate of the true value of θ; we want to fit the model to the data. As Frequentists, we could approach this is many ways, but a good one would be to find the value of θ which yields the maximum likelihood, $p(x|\theta)$. This maximum likelihood estimate (MLE) will be a consistent and unbiased estimate of the true θ under some very general assumptions (Cramer, 1946). And if we wanted a confidence interval for this parameter, we might be able to use the likelihood function to derive this interval. However, these methods can become unreliable as we seek more complex models with many interacting parameters. In Chapter 2, I give a technical discussion of the shortcomings of MLE-based methods for modeling biophysical systems. An inevitable trend is toward biophysical models of higher complexity. In these settings, common inference methods (MLE) are insufficient to assure us that our models are useful and accurate. The Bayesian approach, with the aid of MCMC, allows us to use rigorous inference methods with arbitrarily complex models and gives us a sophisticated way of evaluating models and biophysical parameters.

1.3 Chapter Overview

In the following chapters, I describe the benefits gained by incorporating Bayesian methods into the quantitative modeling of biophysical systems.

In Chapter 2, I describe the pitfall of parameter non-identifiability. Models of biophysical systems are often evaluated on their ability to explain controlled experiments and estimates of relevant biophysical parameters are those which provide the best *fit* to the data. However, common models often involve a large number of non-independent parameters, the result being that (potentially infinitely) many combinations of the model parameters could all yield *identical* data. In these cases, the parameters are not *identifiable*, since a good fit to the data provides no guarantees that the estimated parameter values are close to the true values. I discuss the shortcomings of MLE-based approaches for parameter inference especially with respect to their ability to detect and diagnose identifiability. I then demonstrate that a Bayesian approach to parameter estimation resolves this issue; by efficiently sampling posterior distributions, we gain a direct diagnostic of identifiability. This approach is demonstrated with several relevant applications including ligand binding curves and dynamical systems.

In Chapter 3, I provide a deeper look at the scourge of parameter identifiability by focusing on an emerging biophysical technique called single molecule photobleaching. This experimental technique is useful for determining the stoichiometry of protein complexes. The data derived from the methods are draws from a stochastic process and are used to inform conclusions regarding arrangements and interactions of proteins. In this chapter, I demonstrate that the analysis and interpretation of this data is non-trivial, since the underlying inference model can be nonidentifiable in certain experimental regimes. To overcome these uncertainties, I lay out a rigorous probability model for inference and parameter confidence when using this technique.

In Chapter 4, I describe the use of nonparametric Bayesian inference, which provides a flexible class of infinite dimensional probability distributions, allowing us to circumvent the problems of model selection. In effect, we can use nonparametric Bayes methods in order to *extract structure* from data instead of assuming models beforehand. After describing theoretical aspects of nonparametric Bayes, I demonstrate its utility with several diverse applications in single molecule biophysics. I show that using a Dirichlet process mixture model, we can analyze single ion channel dwell-times in order to discover the number of biophysical states hidden in the data. I then show that a hierarchical Dirichlet process hidden Markov model can be used to analyze a variety of single molecule time series, with applications to electrophysiological recordings, single molecule photobleaching, and single molecule FRET. Finally, I show that with a hierarchical Dirichlet process aggregated Markov model, we can analyze single ion channel time series without assuming a model, and we can discover the hidden open and closed states in the data. These novel methods promise to bring a new level of rigor and power to the study of single molecule biophysics.

In the Appendix, I provide an accessible introduction and tutorial on Bayesian inference. Readers who are entirely unfamiliar with Bayesian methods may benefit from reading this appendix early on, as the description of the methods is somewhat terse and technical within each chapter.

Chapter 2

Determination of Parameter Identifiability in Nonlinear Biophysical Models: A Bayesian Approach

The majority of the text and figures presented here have been published previously in the *Journal of General Physiology*:

Hines, K.E., T.R. Middendorf and R.W. Aldrich (2014). Determination of Parameter Identifiability in Nonlinear Biophysical Models: A Bayesian Approach. Journal of General Physiology. 143(3): 401-416.

Co-author contributions: T.R. Middendorf assisted with developing these results and R.W. Aldrich supervised the project.

Abstract A major goal of biophysics is to understand the physical mechanisms of biological molecules and systems. Mechanistic models are evaluated based on their ability to explain carefully controlled experiments. By fitting models to data, biophysical parameters that cannot be measured directly can be estimated from experimentation. However, it might be the case that many different combinations of model parameters can explain the observations equally well. In these cases, the model parameters are not identifiable: the experimentation has not provided sufficient constraining power to enable

unique estimation of their true values. We demonstrate that this pitfall is present even in simple biophysical models. We investigate the underlying causes of parameter non-identifiability and discuss straightforward methods for determining when parameters of simple models can be inferred accurately. However, for models of even modest complexity, more general tools are required to diagnose parameter non-identifiability. We present a method based in Bayesian inference that can be used to establish the reliability of parameter estimates, as well as yield accurate quantification of parameter confidence.

2.1 Introduction

A major goal of biophysics is to understand the physical mechanisms of biological molecules and systems. The general approach to rigorously evaluate mechanistic hypotheses involves comparison of measured data from well-controlled experiments to the predictions of quantitative physical models. A candidate model (and the mechanism which it implies) is rejected if it does not quantitatively fit all available data. For models that agree with the data, the fits provide estimates for the model parameters, which represent system properties of interest that cannot be measured directly, such as binding affinities, cooperative interactions, kinetic rate constants, etc. An extensive literature exists concerning methods for finding sets of parameter values that provide the best fit of model to data (Jennrich and Ralston, 1979; Johnson and Faunt, 1992; Johnson, 2010), but the important issue of determining and characterizing the confidence of parameter estimates in models with several, usually non-independent, parameters is more challenging and less well developed. Here, by way of examples from fairly simple and common biophysical models, we consider two related issues: i) for a given model and a given type of data, are the parameters of a model uniquely constrained by the measurements? and ii) how confident can one be in the parameter values obtained by fitting a model to data?

To illustrate the issues of confidence in parameter estimation we consider ligand activation of the macromolecular receptor calmodulin (CaM). CaM plays a central role in many biological signaling processes and has been studied extensively (Cheung et al. 1978; Cheung 1980; Hoeflich and Ikura 2002). This protein has four non-identical EF-hand binding sites for calcium ions. Upon activation by calcium, CaM can interact with over 300 known effector proteins (Crivici and Ikura, 1995; Yap et al. 2000). Calcium binding data for CaM are commonly analyzed using a four-site sequential binding model (Figure 2.1A, top), in which the ligand binding events are quantified by the macroscopic equilibrium constants K_1, K_2, K_3 and K_4 for the four binding steps. Mechanistically these four parameters reflect intrinsic binding affinities and potential cooperative interactions between the sites, as originally proposed by Adair to describe cooperative oxygen binding to hemoglobin (Adair, 1925). Figure 2.1A, bottom, adapted from Stefan et al., 2009, shows calcium binding curves from several studies (Crouch and Klee, 1980; Porumb, 1994; Peersen et al, 1997); Figure 2.1B shows the corresponding estimates for parameters K_1 through K_4 reported by these groups, and those from three related studies

(Burger et al., 1984, Linse et al., 1991, Haeich et al., 1981). Strikingly, while the data from these groups are in good agreement, the parameter estimates differ significantly: for some parameters, the reported estimates vary more than 25-fold.

What underlies this large uncertainty in binding parameter estimates? The problem may be intrinsic to data fitting, such that a given binding curve is fit arbitrarily well by many combinations of parameter values, regardless of data quality. Alternatively, it could be a consequence of noise in the data, in which case more precise experimentation may place tighter constraints on the parameter values. We investigated these possibilities using simulated data. A synthetic binding curve with no added noise was generated using the binding parameter estimates from a single study (Linse et al., 1991) (smooth curve in Figure 2.1C). Systematic exploration of the parameter space of the sequential binding model revealed that no single set of parameter values provides a clear best fit to the synthetic data. Rather, many parameter sets, covering a wide range of values for each parameter, fit the data with less than one percent RMS deviation. (This value is a conservative threshold for identifying excellent fits, since real binding measurements are unlikely to surpass this noise criterion). Two representative excellent fits are shown in Figure 2.1C, and the corresponding parameter values for these fits are shown in Figure 2.1D. Note that the scale of the vertical axis in Figure 2.1D spans nine orders of magnitude, indicating that parameter sets with very different apparent affinities yield similar binding curves. For the set of points shown as squares in Figure

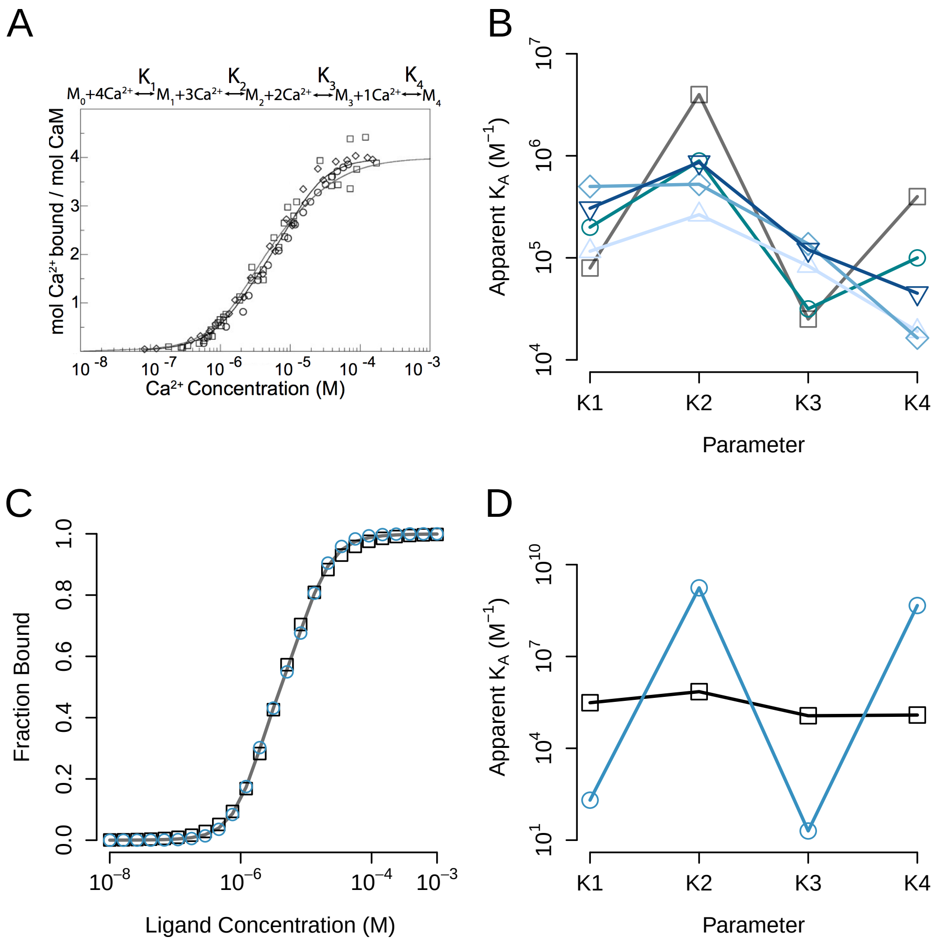

Figure 2.1: Estimating the parameters of a four-site binding model. (A) Calmodulin binding data from multiple experimental groups (adapted from (Stefan et al., 2009)). Mi refers to CaM with i bound calcium ions. (B) Parameters obtained by fitting the sequential binding model in (A) to experimental data from five groups. Parameters are from Linse et al., 1991 (squares), Haeich et al., 1981 (rhombi), Porumb, 1994 (circles), Burger et al., 1984 (triangles), Crouch and Klee, 1980 (inverted triangles). (C) Synthetic, noiseless binding data (solid curve) calculated using parameters from (Linse et al., 1991). (D) Two distinct parameter sets that yield excellent fits to the data. For parameters shown as squares, all binding sites have nearly identical affinity, consistent with weak cooperativity. The parameters shown as circles are consistent with strong cooperative interactions between the binding sites. The binding curves generated from these parameters are plotted in C as circles and squares.

2.1D, the equilibrium constants K_1-K_4 are nearly equal. For the parameter set shown as circles in Figure 2.1D, the apparent affinity for each binding event is very different, consistent with large differences in binding affinity or strong cooperative interactions between the sites. Comparison of the two fits shown in Figure 2.1C indicates that even high-quality binding data (1 percent RMS noise) lack the power to distinguish between mechanisms with very different binding affinity and/or cooperativity. When fitting data, not only the best fit, but also the uniqueness of the fit must be determined to understand the confidence one can have in the estimates. Otherwise, the uncertainties in model parameter values may be so large (as in Figure 2.1D) that they preclude even qualitative insights into the mechanism of the process.

If fitting a model to data is to yield accurate and meaningful parameter estimates, the complexity of the model must be commensurate with the constraining power of the data. We show that this requirement is often not met when typical models are used to analyze common types of experimental data such as binding curves and kinetic time series. We present multiple methods for assessing the uniqueness of parameter estimates obtained from fitting experimental data. For some simple models that allow analytical solutions, we describe a method for determining the maximum number of parameters that can be meaningfully quantified by regression analysis. This work builds upon previous investigations (Reich, 1974; Astrom and Bellman 1970; Straume and Johnson 2010), and such methods are effective for simple models. However, any biophysically realistic models of protein signaling and conformational change, such as ion channel gating, will not be tractable. Therefore, a general method is needed for estimating parameters accurately regardless of model complexity. To this end, we present a framework based in Bayesian inference that employs Markov chain Monte Carlo (MCMC) sampling. By providing distributions of parameter values consistent with the available data, this method yields accurate estimates of model parameters (and their uncertainties) and can be used to determine whether those estimates are unique. In addition, the method possesses significant computational advantages over approaches that sample error surfaces exhaustively.

2.2 Results

In the following sections, we explore the reliability of parameter estimates obtained by fitting various models to common types of biophysical data. Experimental data is generally fitted using regression methods. The theory for nonlinear regression assumes that there is a point in the parameter space that yields a minimum local (though hopefully global) error between model and data. Furthermore, it is assumed that the error contours surrounding this point are well approximated by ellipsoids (this geometry stems from a quadratic error function) (Seber and Wild, 2003; Seber and Lee, 2003). When these conditions hold, we have strong statistical guarantees that the true parameter value is located within some bounded interval from our estimated value. A variety of optimization methods can then be used to obtain a best point estimate for the parameters and to construct confidence intervals defining the uncertainty in the parameters. For many models of biophysical interest, however, multiple parameter combinations (often with very different values) produce nearly identical observables. In these cases, the assumptions of nonlinear regression theory do not hold, and we lose the statistical assurance that our point estimates of the parameters are close to the true values. We next employ simple example models to investigate when such issues arise and to demonstrate that this problem may occur frequently in biophysical investigations.

2.2.1 Two-site, three-parameter binding model: a case of structural non-identifiability

Consider a cooperative model of ligand binding to a receptor with two inequivalent binding sites (Figure 2.2A). The microscopic association equilibrium constants of the sites are denoted K_I and K_{II}, and an additional cooperativity parameter (F) quantifies the degree to which binding events at one site can enhance (or hinder) binding at the other site. Note there is only one free parameter for cooperativity, due to the detailed balance constraint. A simulated binding curve is shown in Figure 2.2B which was generated with parameters $\{K_I, K_{II}, F\}$ = $\{500\ \mu\mathrm{M}^{-1}$, $500\ \mu\mathrm{M}^{-1}$, $1\}$. A second binding curve is also shown, generated with parameters $\{K_I, K_{II}, F\}$ = $\{997.5\ \mu\mathrm{M}^{-1}$, $2.5\ \mu\mathrm{M}^{-1}$, $100\}$. Though these curves were generated from very different parameter values, they overlay exactly. How is it that multiple points in the parameter space of this model can yield identical binding curves?

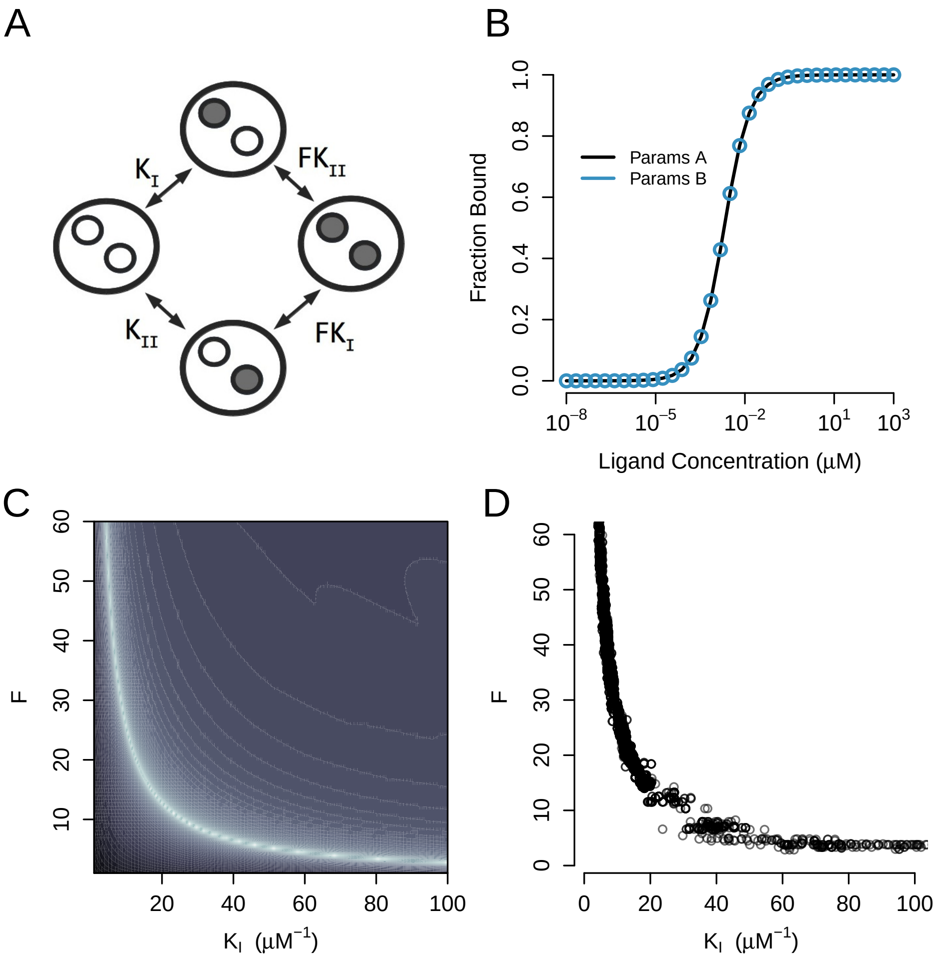

Figure 2.2: Parameter estimation for a two-site cooperative binding model. (A) Diagram of a model which assumes two binding sites with microscopic association constants KI and KII, and cooperativity factor F. (B) Two parameter sets with very different values yield identical simulated binding curves. Parameter set A: $\{KI, KII, F\} = \{500\ \mu M^{-1}, 500\ \mu M^{-1}, 1\}$; Parameter set B: $\{KI, KII, F\} = \{997.5\ \mu M^{-1}, 2.5\ \mu M^{-1}, 100\}$. (C) Log-error surface of a region of the F-KII parameter space. The curve generated from Parameters A was used as a reference curve. Binding curves were calculated for various points in the parameter space, and the total error between the two curves was computed. Areas of lighter shading correspond to areas of less error. (D) MCMC samples drawn from the joint posterior distribution of KI and F. The curved structure of the posterior distribution indicates that the model parameters are not identifiable using this binding data.

We used the smooth curve in Figure 2.2B as simulated data (without added noise) and explored the values of parameters F and KI systematically (Figure 2.2C). For every point in this grid, the values of F and KI were fixed and the third parameter (K_{II}) was varied to generate the best fit to the reference curve using nonlinear least-squares regression (Levenberg, 1944; Marquardt, 1963). The residual sum-squared-error between the model and the data was then determined for each F, K_I pair. This error surface is represented in Figure 2.2C, with lighter shading corresponding to lower total error. No single combination of parameter values in this surface provides a best fit to the reference curve. Rather, a vast contour through the parameter space yields equally good fits to the synthetic data. The minimum error contour (lightest color in Figure 2.2C) extends infinitely in both directions of the parameter space, even as the total error approaches zero. The shape of this contour illustrates how the parameter values compensate systematically over wide ranges to fit the data. (Note that the true value of K_I (500 μM^{-1}) is far to the right of the plot shown at this scale.) An experimental scientist confronted with the error surface in Figure 2.2C would reach several discouraging conclusions: i) finding a good fit of the cooperative model to typical binding data provides no guarantee that the inferred parameter values are close to the correct values; ii) the range of parameter values consistent with an exact fit to an experimental binding curve is infinite; and iii) more careful experimentation to reduce the noise in the data will not improve matters.

Situations in which fitting a model to data does not yield unique and optimal parameter estimates is well-documented in the control theory literature (Bellman and Astrom, 1970; Cobelli and DiStefano III, 1980; Walter and Pronzato, 1997). The parameters of the model in Figure 2.2 are not structurally identifiable: there is not enough constraining power, even in noiseless data, to enable a unique estimate of their true values. It is easy to imagine that correlations between the numerous parameters in complex models could yield non-unique fits to data. However, the results in Figure 2.2 illustrate that interactions between the three parameters in a very simple model can be so effective that fitting of high quality data provides little meaningful constraint on the parameter values. How can we determine whether the parameters of a model are structurally identifiable when constrained by a measurement?

2.2.2 Rank-deficient regression

When the experimental observables of a system can be expressed analytically in terms of the model parameters, the data fitting problem can be cast in closed form, and simple methods can be used to test for parameter identifiability (Seber and Lee, 2003). Returning to the model of Figure 2.2A, we define a parameter vector a with components $a_1 = K_I$, $a_2 = K_{II}$, and $a_3 = FK_IK_{II}$. If our observable signal (y) is the fraction of sites occupied by ligand at concentration x, then

$$y(x) = \frac{a_1x + a_2x + 2a_3x^2}{2(1 + a_1x + a_2x + a_3x^2)}. \tag{2.1}$$

If we can cast the observable as a linear system of the model parameters, then we can simply use linear regression for parameter inference. Multiplying both sides of (1) by the denominator of the right hand side linearizes the parameters

$$2y + 2a_1xy + 2a_2xy + 2a_3x^2y = a_1x + a_2x + 2a_3x^2. \tag{2.2}$$

We proceed to find $\hat{a}$, an optimal estimate of the parameters, by minimizing the error between the model and the data. For our cost function, S, we use the sum of the squared error between observations y_i and model predictions $y(x_i)$:

$$S = \sum_{i=1}^{n} (y_i - y(x_i))^2 . \tag{2.3}$$

Substituting in the expression for $y(x_i)$ yields:

$$S = \sum_{i=1}^{n} \left(a_1x_i(2y_i - 1) + a_2x_i(2y_i - 1) + 2a_3x_i^2(y_i - 1) + 2y_i\right)^2 . \tag{2.4}$$

The optimal estimate of the parameters is obtained by minimizing S. The partial derivatives of S with respect to the components of the parameter vector a are:

$$\begin{aligned}
\frac{\partial S}{\partial a_1} &= \sum_{i=1}^{n}(a_1(4x_iy_i - 2x_i)^2 + a_2(4x_iy_i - x_i)^2 + a_3(4x_iy_i - 2x_i)(2x_i^2y_i - 2x_i2) \\
&\qquad +4y_i(2x_iy_i - x_i)) \\
\frac{\partial S}{\partial a_2} &= \sum_{i=1}^{n}(a_1(4x_iy_i - 2x_i)^2 + a_2(4x_iy_i - x_i)^2 + a_3(4x_iy_i - 2x_i)(2x_i^2y_i - 2x_i^2) \\
&\qquad +4y_i(2x_iy_i - x_i)) \\
\frac{\partial S}{\partial a_3} &= \sum_{i=1}^{n}(a_1(4x_iy_i - 2x_i)(2x_i^2y_i - 2x_i^2) + a_2(4x_iy_i - 2x_i(2x_i^2y_i - 2x_i^2)) \\
&\qquad +a_3(2(2x_i^2y_i - 2x_i^2)) + 4y_i(2x_iy_i - x_i)).
\end{aligned}$$

Setting these equal to zero yields the estimate $\hat{a}$ and results in a linear system in $\vec{a}$,

$$\mathbf{X}\hat{a} = \vec{R}, \tag{2.5}$$

where $\hat{a}^T = \{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$ and $R^T = \{\sum_{i=1}^{n} 4y_i(2x_iy_i - x_i), \sum_{i=1}^{n} 4y_i(2x_iy_i - x_i), \sum_{i=1}^{n} 4y_i(2x_iy_i - x_i)\}$. The design matrix $\mathbf{X}$ is

$$\begin{pmatrix} \sum(4x_iy_i - 2x_i)^2 & \sum(4x_iy_i - 2x_i)^2 & \sum(4x_iy_i - 2x_i)(2x_i^2y_i - 2x_i) \\ \sum(4x_iy_i - 2x_i)^2 & \sum(4x_iy_i - 2x_i)^2 & \sum(4x_iy_i - 2x_i)(2x_i^2y_i - 2x_i^2) \\ \sum(4x_iy_i - 2x_i)(2x_i^2y_i - 2x_i) & \sum(4x_iy_i - 2x_i)(2x_i^2y_i - 2x_i) & \sum(4x_i^2y_i - 2x_i^2)^2 \end{pmatrix}. \tag{2.6}$$

The solution of this linear system should yield $\hat{a}$, the optimal point estimate of the parameters given the data. However, the first and second columns of the design matrix, (equation 2.6), are identical (and therefore linearly dependent). In general, the rank of a matrix is the number of linearly independent rows (or columns) and a matrix whose rank is less than the total number of rows or columns is called rank-deficient. A rank-deficient matrix defines a linear system with an infinite number of solutions (Seber and Lee, 2003). In this example model, the design matrix is rank-deficient and therefore equation 2.6 specifies not a point estimate, but rather all combinations of the parameters that are optimal fits to the data. By indicating that a unique estimate of the model parameters is not possible, this test of the design matrix for rank-deficiency is effectively a structural identifiability test. This method of assessing parameter identifiability is similar to other proposed methods that employ sensitivity matrices or Fisher information (Cobelli and DiStefano III, 1980), and can be generally applied to any model where the observable can be expressed as a linear system of the model parameters.

2.2.3 Two-site, two-parameter binding model: a case of practical non-identifiability

Demonstrating that a systems design matrix is full-rank is a necessary, but not sufficient, condition for ensuring that the models parameters can be uniquely estimated from experimental data (Jacquez and Greif, 1985; Faller et al., 2003; Raue et al., 2009). Consider the model depicted in Figure 2.3A for a two-site receptor. This sequential binding model (which is a reduced version of the four-site model in Figure 2.1A) assumes that the two singly-occupied binding configurations are equivalent, and has only two macroscopic affinity parameters, which quantify the first and second binding steps. It is straightforward to show that the design matrix for the model in Figure 2.3A is full-rank. Therefore, we might be tempted to conclude that the parameters of this model can be inferred uniquely from binding data.

Figure 2.3B shows two simulated data sets that were calculated using the model in Figure 2.3A with distinct parameterizations. For parameter set A, K1 = K2 = 200 μM^{-1}. For parameter set B, K1 = 100 μM^{-1}, K2 = 1000 μM^{-1}. To one of the curves (shown as circles), Gaussian noise of 2.5 percent variance has been added to mimic the variability of experimental data. Though the curves were generated from very different parameter values, they produce similar curves (apart from the added noise). The error surface (Figure 2.3C) was computed as the difference between a noiseless reference curve (the solid curve in panel B) and the model predictions for a large region of the parameter space of the model. The error contours are curved, and thus, as for the models in Figs. 2.1 and 2.2, parameter compensations can occur so that disparate parameter values yield the same error. However, unlike those in Figure 2.2C, the error contours in Figure 2.3C are bounded, with the lowest-error contours approaching perfect ellipses. Thus, for data with infinite signal-to-noise ratio (no added noise), the two-site, two-parameter sequential model is uniquely identifiable, and fitting of such data would yield accurate parameter estimates. However, the inevitable presence of experimental noise in real data (even at low levels of 2.5 percent variance) would prevent a unique determination of the parameters of this model. Previous authors have documented this phenomenon (Jacquez and Greif, 1985; Vajda et al., 1989; Raue et al., 2009) and distinguish between structural non-identifiability (as in Figure 2.2), in which even noiseless data cannot yield unique parameter estimates, and practical non-identifiability

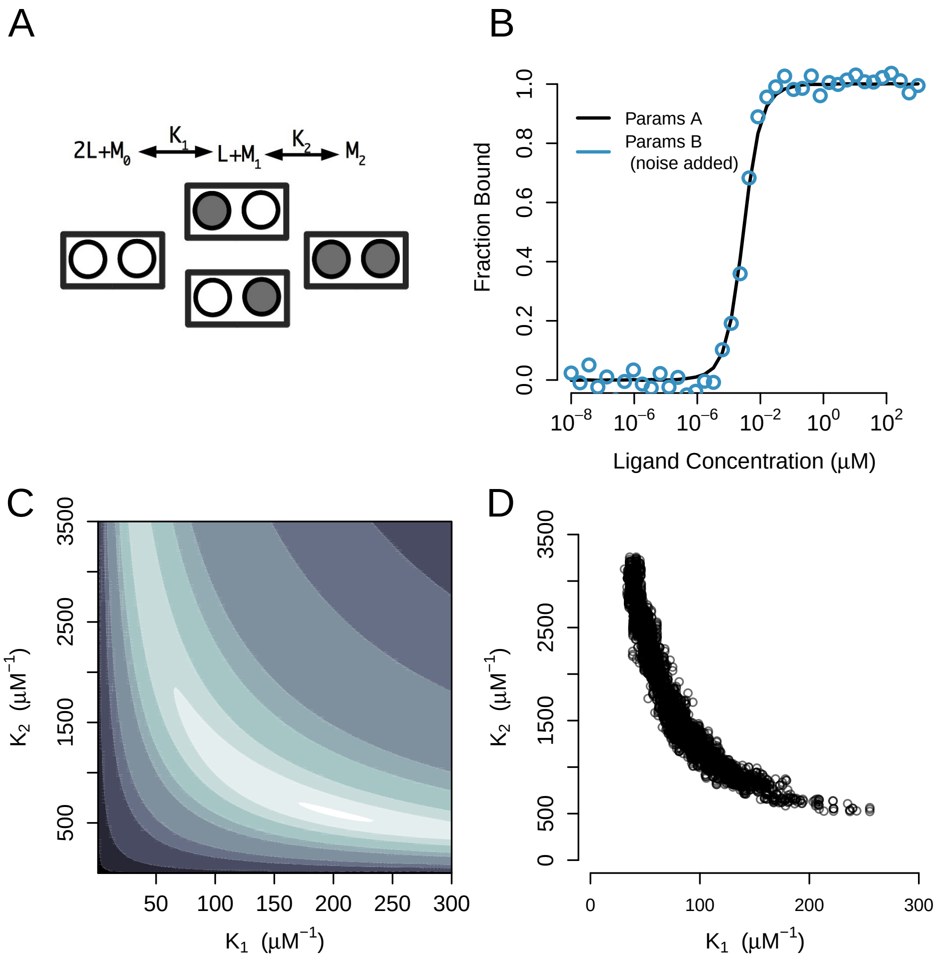

Figure 2.3: Parameter estimation for a two-site sequential binding model. (A) Diagram of model in which macroscopic binding constants K1 and K2 quantify the affinities of the first and second binding steps, respectively. (B) Simulated binding curves for two different parameter sets. Parameter set A is consistent with weak cooperativity between the binding sites (K1 = K2 = 200 μM^{-1}). Parameter set B is consistent with strong binding cooperativity (K1 = 100 μM^{-1}, K2 = 1000 μM^{-1}). Gaussian noise was added to the curve for parameter set B to mimic experimental variability. Though these parameter sets (and their mechanistic interpretations) are quite different, they produce nearly identical observables. (C) Log-error surface in K_1 - K_2 parameter space with respect to the noiseless data curve in B. Large ranges of parameter values produce very similar binding curves. (D) MCMC samples of joint posterior distribution of the parameters when constrained by the noisy curve in B.

(as in Figure 2.3), in which the parameters of a model are identifiable only if data is available with sufficient signal-to-noise.

These examples demonstrate the dangers one incurs when only point estimates or best-fits are considered. It is of vital importance to establish not only the best fit to the data, but the full range of parameters that yield good fits. The typical approach when fitting data to nonlinear models relies on maximum likelihood (ML) theory to estimate parameters (Cramer, 1946). The maximum likelihood estimate (MLE) of a parameter is the point in parameter space that yields the optimum of the likelihood of observing the data given a particular value of the parameters. Asymptotically, the MLE is an efficient and unbiased estimate of a parameter. As an example, if the data are assumed to be normally distributed, then minimizing the sum-squared-error between model and data provides the MLE (Seber and Wild, 2003). Once an optimal point estimate is found, ML theory prescribes that confidence regions can be calculated by identifying the range of parameters that yield likelihoods consistent with the noise in the data (see Colquhoun et al., 2003 for a description of properties of ML estimators for common biophysical systems). For low dimensional models, a grid of the entire likelihood surface might be computed (as in Figs. 2.2 and 2.3C), but this becomes infeasible for larger models (see Discussion). Due to this constraint, it is typically assumed that the likelihood surface is approximately quadratic around the MLE. Therefore, efficient algorithms can be used to identify the MLE and to numerically approximate the local curvature of the likelihood in order to construct a confidence ellipsoid around the MLE. In cases of non-identifiability, this elliptical approximation around the MLE can be quite inaccurate. Take, for example, the likelihood surface of the two parameter binding model when constrained by the noiseless binding curve in Figure 2.3C (K1 = 100 μM^{-1}, K2 = 1000 μM^{-1}). The maximum likelihood estimate would indeed be the true parameter values, and the lowest error contour surrounding this point would be well approximated by an ellipse (lightest color contour in C). However, in the face of realistic experimental noise, our estimate of parameter confidence must take into account all parameter combinations that are consistent with any particular level of error. Such error contours are no longer elliptical, but are curved (due to practical non-identifiability). If we use a quadratic likelihood approximation around the MLE in Figure 2.3C, we would vastly underestimate our parameter uncertainty. It might be the case that the likelihood surface near the MLE is relatively flat in the direction of one (or more) of the parameters and this would suggest non-identifiability. However, we will be unable to distinguish between structural and practical non-identifiability without directly assessing all the regions of parameter space that yield high likelihood. Since exhaustive exploration of parameter space will not be feasible for most realistic models (see Discussion), we next describe a computationally efficient method of exploring parameter spaces.

2.2.4 Bayesian Inference

The previous section demonstrated that in the face of realistic experimental noise, analysis of the design matrix does not provide a sufficient condition for establishing whether the parameters of a model are uniquely constrained by the available data. Therefore, a more general method is needed for determining whether the parameters of a model are both structurally and practically identifiable. In Figs. 2.2C and 2.3C, we explored the entirety of parameter spaces to identify which regions led to low error between model predictions and data. If the parameter values in best agreement with the data are confined to a small region of the parameter space, then the model parameters are identifiable. This approach moves us away from the idea of accepting a single best fit to the data, and instead identifies all regions of the parameter space that are in agreement with the observations. In the language of Bayesian inference, what we seek is called the posterior distribution of the parameters: a probability distribution on the parameter space that assigns higher probability to areas that are in better agreement with the observations. In the following, we demonstrate that a Bayesian approach provides accurate estimates of model parameters and their uncertainty and provides a direct and general method of diagnosing parameter identifiability.

Assume that we have gathered N observations, y_N, in order to infer the true values of m parameters $\{\theta_1, \theta_2, ..., \theta_m\}$, comprising the vector θ. In Bayesian terms, we seek to know $p(\theta|y_N)$, the posterior probability distribution of the parameters, which is the probability (over the entire parameter space) of a particular value of θ having given rise to the observations y_N. To estimate this distribution, we apply Bayes rule,

$$p(\theta|y_N) \propto p(y_N|\theta)p(\theta) \tag{2.7}$$

which states that the posterior distribution of the parameters is proportional to the likelihood of observing the data, $p(y_N|\theta)$, multiplied by the prior distribution of the parameters, $p(\theta)$. If the observations y_i are independent, then the total posterior probability is the product of the posterior probability of each observation,

$$p(\theta|y_N) \propto \prod_{i=1}^{N} p(y_i|\theta)p(\theta). \tag{2.8}$$

For a particular model and some observed data, it is straightforward to compute $p(\theta|y_N)$. The structure of the posterior distribution indicates whether the region of highest posterior probability (the best fits) is localized or extends over a significant fraction of the parameter space, and is thus an indicator of parameter identifiability.

As with the direct calculation of error surfaces (Figs. 2.2C and 2.3C), computing posterior distributions over an entire parameter space by brute force is possible for low-dimensional problems, but quickly becomes infeasible for even modestly sized models. Fortunately, posterior distributions can be computed efficiently using an existing numerical method from the statistics literature called Markov chain Monte Carlo (MCMC) sampling. The theory of MCMC is described elsewhere (Robert and Casella, 2010; Tierney, 1994) and we provide only a brief description. Consider a high-dimensional system for which the brute force computation of posterior probabilities over the entire parameter space is not practical. If we can draw a finite number of independent and identically distributed (iid) samples from the corresponding posterior distribution, then the properties of this finite sample will approximate the properties of the posterior. To generate these iid samples, we simulate a Markov chain whose limiting distribution is the posterior distribution of interest. Generating a Markov chain with a desired limiting distribution can be achieved by a number of methods. Here we rely on one of the simplest-the Metropolis Random Walk algorithm (Metropolis et al., 1953). For iteration i of the algorithm, the Markov chain is in position θ_i in the parameter space. A potential transition to a new point, $\tilde{\theta}$, is generated by a random walk according to the following rules. The posterior probability of this potential point, $p(\tilde{\theta}|y_N)$, is computed and compared to the posterior probability of the current position of the chain, $p(\theta_i|y_N)$. If the proposal point has greater posterior probability than the current point, then it is accepted as a sample from the posterior distribution. If it has lower posterior probability, then it is rejected with probability proportional to the decrease in posterior probability. Thus, transitions of the Markov chain are accepted with probability α, where

$$\alpha = min\left(1, \frac{p(\tilde{\theta}|y_N)}{p(\theta_i|y_N)}\right). \tag{2.9}$$

This rule allows the chain to move efficiently toward areas of high posterior probability but also provides a mechanism to move away from local minima in the posterior distribution by allowing transitions to regions of lower posterior probability. The Markov chain produced by this algorithm explores the parameter space in proportion to the posterior probability and provides a finite number of iid samples from the posterior distribution. This method can be used to efficiently approximate posterior distributions of arbitrarily high dimension. The following section illustrates the use of MCMC to assess parameter identifiability for some common biophysical models.

2.2.5 Applications

A common form of parameter inference involves fitting a candidate model to observations drawn from a controlled experiment. Though the Bayesian framework presented here is general, we focus primarily on curve-fitting applications due to their prevalence in experimental science. We assume that each observation, y_i, is a function of some independent variables, x_i, and that the model of interest defines the function, $f(x_i, \theta)$, which depends on the model parameters, θ. We seek to identify the values of θ that lead to the best agreement between y_N and $f(x_N, \theta)$.

Our probability model considers that each observation y_i is the result of $f(x_i, \theta)$ plus some experimental noise, which is assumed to be normally distributed (although this assumption is not necessary). Each observation is drawn from a normal distribution $N(m, \sigma^2)$ whose mean, m, is equal to the model prediction, $f(x_i, \theta)$, for some particular values of the parameters, and whose variance, σ^2, is due to noise of any kind:

$$y_i \sim N(f(x_i, \theta), \sigma^2). \tag{2.10}$$

The posterior probability distribution then becomes,

$$p(\theta | y_N) \propto \prod_i N(y_i | f(x_i, \theta), \sigma^2) p(\theta). \tag{2.11}$$

In the following applications, the prior distribution, $p(\theta)$, is a flat distribution (a truncated uniform distribution). While this form of the prior works well for the simulated datasets we use for illustration, it is in general an improper prior and more thoughtful prior distributions should be used in practice. By using MCMC, we generate a Markov chain that preferentially explores regions of the parameter space that lead to high posterior probability (i.e., the best fits to the data).

2.2.6 Binding Models

We showed earlier, using algebraic techniques, that the three-parameter binding model of Figure 2.2A is not structurally identifiable. Consistent with this finding, the error surface (Figure 2.2C) revealed an unbounded zero-error contour through the parameter space of this model. MCMC samples from the joint posterior distribution of F and K_I (Figure 2.2D) show that this distribution is highly curved, indicating that a large range of values of these parameters is in good agreement with the data. The MCMC approach leads to the same conclusion as the error surface, but with a much-improved computational efficiency and potential for scalability (see Discussion).

A thorough mapping of the error surface for the two parameter model of Figure 2.3A, shown in Figure 2.3C, revealed that this model is not practically identifiable. While unique best fit parameters can be obtained in theory, this is not possible for data with a realistic signal-to-noise ratio. Consistent with this observation, the MCMC approximation to the posterior distribution for this model (Figure 2.3D) revealed that the two parameters of the model can compensate for one another to produce good fits to the data. In this case, the noisy data of Figure 2.3 is used to constrain the model parameters for MCMC. When faced with this level of noise in the data, parameters are able to compensate, as revealed by the curved structure of the posterior distribution. However, in contrast to situations of structural non-identifiability (for which it is impossible to constrain parameter estimates usefully), we would conclude that the true values of the parameter lie within a certain bounded region (by constructing a 95% credible interval), when constrained by this data: parameter K1 is likely between 50 and 250 μM^{-1}, and parameter K2 is likely between 200 and 2000 μM^{-1}. While this level of confidence is a considerable improvement over the situation of Figure 2.2, these large uncertainties may still prevent us from achieving a useful level of mechanistic insight. For example, there are posterior samples corresponding to $\{KI, KII\} = \{100, 1000\}$ and $\{250, 250\}$, each of which is a valid explanation of the data. Therefore, while we can put reasonable bounds on parameter estimates, we may not be able to draw even qualitative conclusions regarding mechanism.

2.2.7 Kinetic Models

Many chemical and biochemical systems can be described by kinetic models (such as in Figs. 2.4A and 2.6A) comprising systems of coupled differential equations. Typical experimental investigations of these systems involve monitoring the time course of the state populations in response to a perturbation to determine the transition rate constants. Numerous methods have been proposed to assess parameter identifiability in these so-called compartmental systems (Cobelli and DiStefano III, 1980; Godfrey et al., 1982), including Laplace transforms (Walter and Pronzato, 1997), and information matrices (Bellman and Astrom, 1970). However, the practical non-identifiability of model parameters for many biological systems may not be detected by matrix methods (Raue et al., 2009). An alternative approach to assess identifiability involves computing low-dimensional error surfaces in the relevant parameter space directly (Johnson et al., 2009b, Johnson et al., 2009a, Raue et al., 2009). In the following, we use the more efficient Bayesian framework to determine whether candidate kinetic models are uniquely constrained by a given observable.

The kinetic scheme of Figure 2.4A comprises three states connected sequentially. Suppose that our observable signal is the population of state B over time (Figure 2.4B) with additional Gaussian noise. The independent variable x_i represents time and the model prediction, $f(x_i, \theta)$, is the solution to the system of differential equations represented by the diagram in Figure 2.4A with the parameter set θ. We use these observations to estimate the posterior distribution of the model parameters by generating 50,000 MCMC samples. The structure of the posterior distribution will indicate whether the four transition rates $\{a, b, r, s\}$ are uniquely constrained by this observable.

At the top of Figure 2.4C, the trajectory of one dimension of the Markov chain is plotted (corresponding to parameter a). The thin trace represents the marginal likelihood throughout the course of the simulation. The marginal likelihood, which quantifies the total goodness of fit between the model and the data, starts at a low value, since the simulation is initialized at an arbitrary point in parameter space that is likely a poor fit to the data. As the Markov chain evolves, the marginal likelihood improves and eventually plateaus after approximately 100 iterations; this initial period is termed the burn-in and these samples are discarded. After this convergence, the Markov chain has reached stationarity and all subsequent transitions provide iid samples from the posterior distribution (see Discussion). The estimate of parameter a (thick trace) moves in large jumps initially but eventually settles near the true value of 15. The trajectories of each of the other parameters are also plotted for

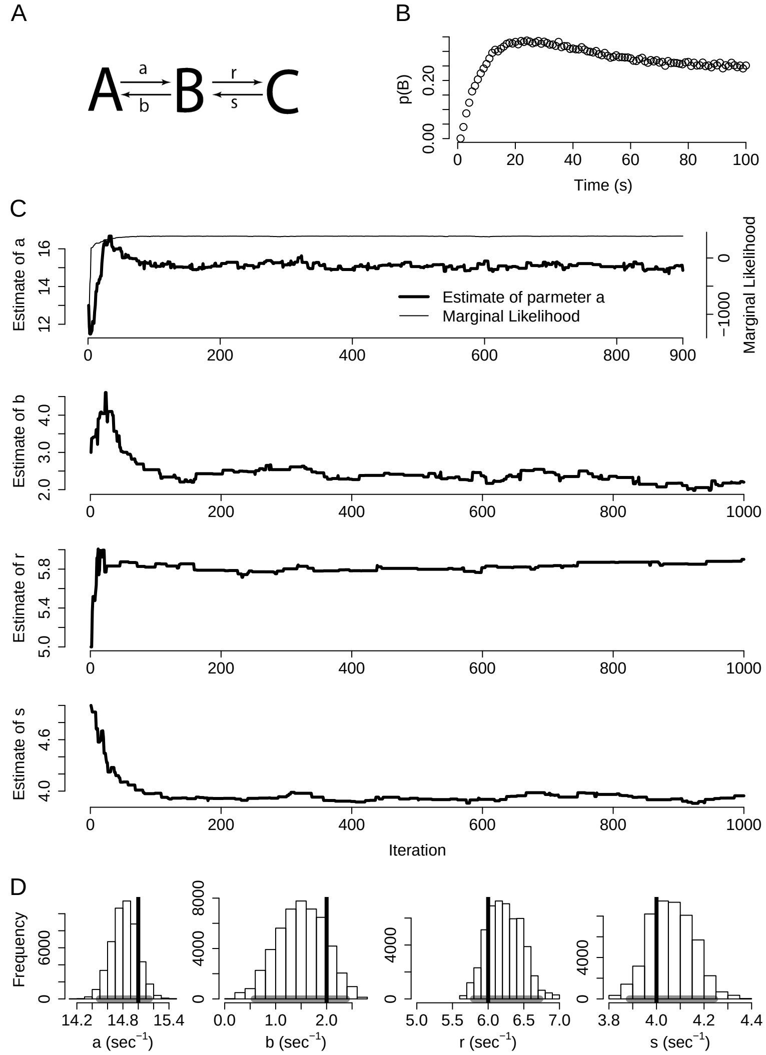

Figure 2.4: Application of MCMC to dynamical systems. (A) A kinetic model with three states and four free parameters. (B) Time course of the population of state B. (C) MCMC results when inferring the parameter values from the data in (B). Top panel shows one dimension of the Markov chain (corresponding to parameter a) throughout the course of the simulation (thick trace). Thin trace is the corresponding marginal likelihood. Lower traces are the corresponding trajectories for the remaining parameters. (D) Histograms of the marginal posterior distribution of each parameter shown along with the true values (vertical lines) and the corresponding 95% credible interval (horizontal line segment).

the first 1,000 iterations in panel C. In each case, the chain explores a small region of the parameter space but does not stray far from the optimal estimate, especially after the Markov chain converges. Each of these transitions represent an iid sample from the posterior distribution and is a valid estimate of the parameter. Therefore, the transitions of the Markov chain, taken in aggregate, approximate the total uncertainty in each parameter (called the marginal posterior distributions). Panel D shows histograms of the estimate of each parameter (after the burn-in) as well as the true values (vertical lines). Such an approximation of the marginal posterior distributions can be used to derive credible intervals for each parameter. While the peaks of the posterior distributions do not all coincide with the true parameter values, 95% credible intervals (horizontal line segments below the histograms) contain the true values.

Since MCMC samples are drawn from the total joint posterior distribution of the parameters, they can be used to assess any pair-wise (or higher order, if desired) correlations between the parameters. Similar to Figures 2.2 and 2.3 D, pair-wise joint posterior distributions are shown in Figure 2.5, using the same MCMC samples from Figure 2.4C. The density of samples is used to generate a map such that areas of lighter shading correspond to areas of higher posterior probability. In this way, the four-dimensional posterior distribution of this model is projected into each two-dimensional subspace. The ensemble of good fits to the data is confined to small regions of the parameter space which contain the true parameter values, and therefore the parameters of this

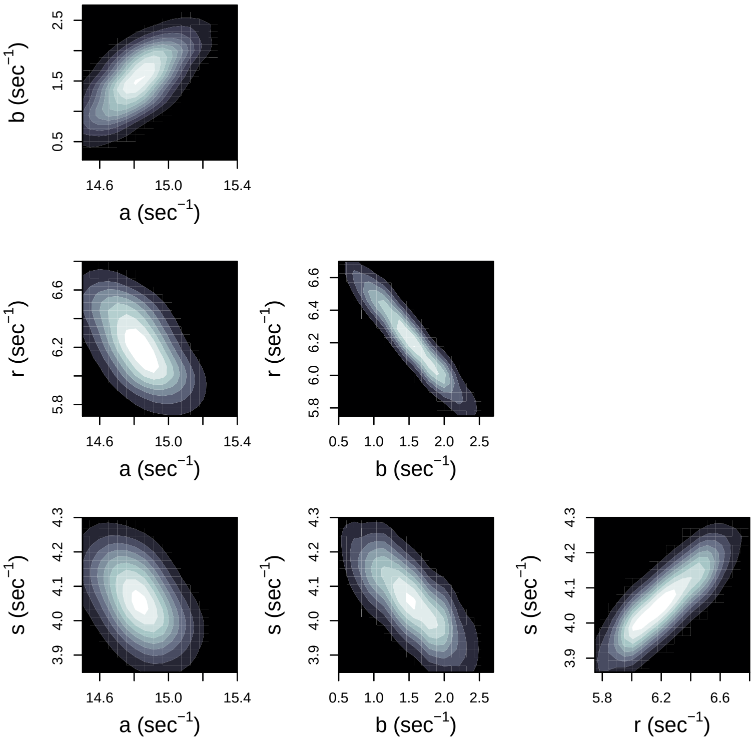

Figure 2.5: MCMC can be used to assess parameter correlations. The MCMC trajectories from Figure 2.4C were used to visualize all the pairwise correlations between model parameters. The density of MCMC samples has been used to generate maps where areas of lighter shading correspond to higher posterior probability. In contrast to Figures 2.2 and 2.3 D, the joint posterior distributions of these model parameters are approximately elliptical, indicating that the optimal estimates of the model parameters are contained in a small, bounded region of the parameter space.

model are identifiable.

As a counter example, consider the more complex model of Figure 2.6A, which has five states and eight free parameters. In this case, assume that the observable is the combined populations of states D and E (Figure 2.6B) with additional Gaussian noise. The panels in Figure 2.6C show 100,000 samples from the resulting MCMC trajectories for each of the eight parameters. At top left, one dimension of the Markov chain (corresponding to parameter a) is shown along with the marginal likelihood (thin trace). The portions of the trajectories plotted in panel C after the marginal likelihood settles represent excellent fits to the data. In nearly every case, a large range of values is sampled, all of which yield comparable marginal likelihood, meaning that they provide excellent fits to the data and can be considered valid estimates. This unbounded exploration of the parameters demonstrates that these parameters are not identifiable when constrained by this data.

2.3 Discussion

2.3.1 Parameter identifiability and model selection

The work described here was motivated by the striking observation that typical binding data place very weak constraints on the magnitudes of affinity parameters for multi-site receptors. We showed that many parameter sets, with affinity values varying by over four orders of magnitude for each of the steps in a sequential binding model for calmodulin, produced binding curves differing by less than 1% RMS (Fig 2.1). Even if binding data could be

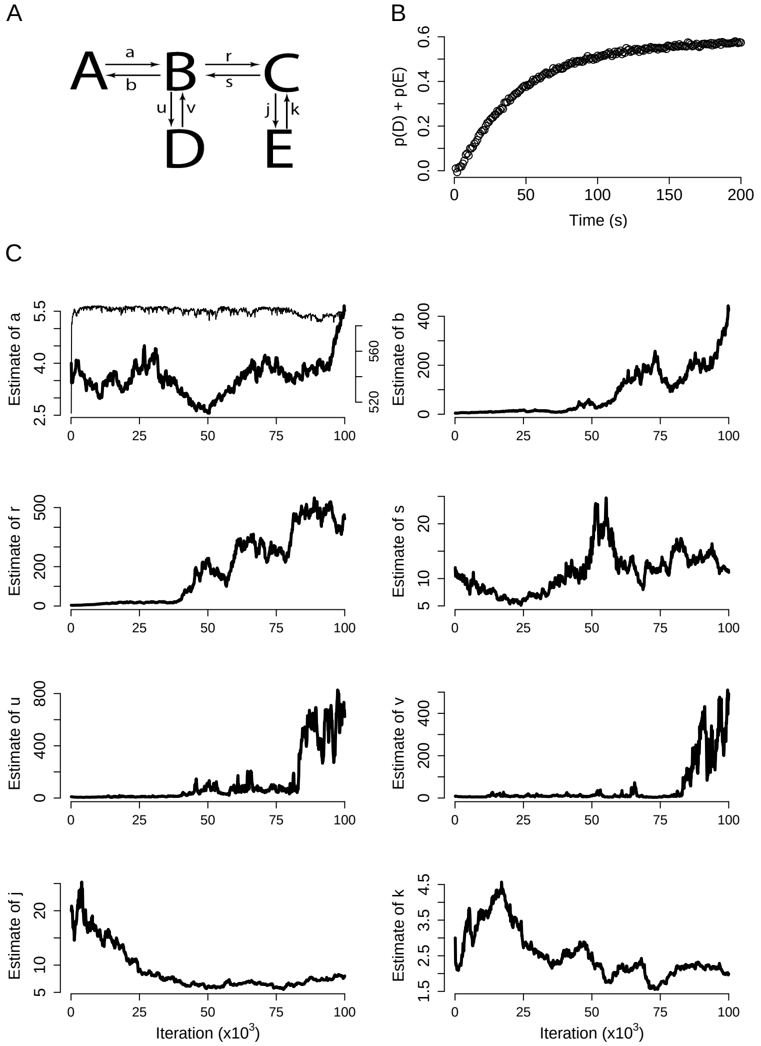

Figure 2.6: MCMC can detect non-identifiable models. (A) A five-state model with eight free parameters. (B) Time course of the combined populations of states D and E with parameters $\{a, b, r, s, u, v, j, k\} = \{3, 3, 5, 10, 9, 9, 20, 4\}$ (values in s^{-1}). (C) Result of using MCMC to infer parameter values. At top left, the thick black trace is one dimension of the Markov chain (corresponding to parameter a) throughout the course of the simulation. The thin trace is the corresponding marginal likelihood. The MCMC trajectories of the other model parameters are also shown. Since the marginal likelihood stabilizes, but most of the parameter estimates do not, this model is not identifiable when constrained by this measurement.

obtained with this low noise level, the enormous uncertainties in the derived parameter estimates severely limit the usefulness of this data for developing a quantitative model for calcium activation of CaM. Parameter estimates are not unique even for simple two-site binding models comprising only two or three parameters (Figs. 2.2 and 2.3). Similar problems affect parameter estimation for dynamical models used to analyze biochemical kinetic data (Figure 2.4). When model parameters are not identifiable, one has little confidence that estimated values are close to the true values. For some model/data combinations, the data are fit arbitrarily well by many combinations of parameter values, and the uncertainties in the model parameter estimates are unbounded, even for noiseless data. These systems are structurally non-identifiable: the model contains more parameters than can be supported even by perfect data (Bellman and Astrom, 1970; Cobelli and DiStefano III, 1980; Walter and Pronzato, 1997). A structurally non-identifiable system is analogous to an under-determined system of algebraic equations, which has an infinite number of solutions.

Structural non-identifiability can often be detected using algebraic methods, as in our demonstration that the design matrix for the two-site cooperative binding model in Figure 2.2 is rank-deficient. Identifiability analysis indicates that this model is over-parameterized: we are attempting to extract three model parameters (F, K_I, and K_{II}) from curve fitting, whereas the rank of the design matrix for this system, which specifies the maximum number of parameters that can be quantified by fitting ideal (i.e., noiseless) total binding data, is two. The parameterization of the cooperative model is designed to address two fundamental questions about a receptor with two binding sites: i) are the site affinities unequal (i.e., is K_I not equal to K_{II}?), and ii) does binding at one site influence binding at the other site (i.e., is F not equal to 1?). Since it requires three parameters to quantify these properties, it is not possible to extract site affinities and cooperative interactions from this single type of experiment. Meaningful regression analysis of these data with this model requires a simpler parametrization than that in equation 2.1, such as

$$y(x) = \frac{b_1 x + 2b_2 x^2}{2(1 + b_1 x + b_2 x^2)}. \tag{2.12}$$

One could then make the simplifying assumption that the binding site affinities are equal, and define the parameters as $\{b_1, b_2\} = \{2K, FK^2\}$, where $K_I = K_{II} = K$. Alternatively, one could assume that the sites do not interact cooperatively, and define the parameters as $\{b_1, b_2\} = \{K_I + K_{II}, K_I K_{II}\}$. If both of these options were deemed unsatisfactory, then other types of data would need to be recorded. A new round of structural identifiability analysis would then indicate whether three parameters could be extracted from fitting the enhanced data set. This example illustrates how structural identifiability analysis can provide an upper limit on what can be learned about a system through experimentation. It is necessary that the parameters of a model are structurally identifiable with respect to a given type of data for inference to even be possible. However, the uncertainties in the parameters estimated by regression analysis of such a system might still be unacceptably large. For these practically non-identifiable cases, the uncertainty in parameter values is linked to the amount of noise in the data, such that meaningful parameter estimates are obtained only if the noise amplitude is sufficiently small (Jacquez and Greif, 1985; Faller et al., 2003; Raue et al., 2009). Establishing that a system is practically non-identifiable is inherently subjective, because the acceptable parameter uncertainty must be weighed against the difficulty (or impossibility) of improving the signal-to-noise ratio of the data to a specified level. A useful approach, which we have followed here, is to determine by simulation the precision in parameter estimates that is required for gaining useful mechanistic insight into the system under study. If this precision requires data with a signal-to-noise ratio that is not achievable in practice, then the parameters are practically non-identifiable.

Using the algebraic approach described here, it is easily shown that the parameters of the four-site sequential model (Figure 2.1A) are structurally identifiable (i.e., the associated design matrix is full-rank). However, the simulations in Figure 2.1 indicate that synthetic binding data with extremely low (1% RMS) noise are not sufficient to constrain the values of the affinity parameters K_1-K_4 to within less than four orders of magnitude. Thus, the four-site model parameters are practically non-identifiable when constrained by this type of data. The large parameter uncertainties prevent even qualitative insights about cooperative interactions in CaM.

For the examples in Figs. 2.1-2.6, we explored the uniqueness of parameter estimation by fitting an assumed model to data. However, in real-world experimental investigations, the correct model is usually not known. There are often several different competing schemes for describing a given biophysical phenomenon, and thus identifying a satisfactory model is an important aspect of the overall modeling process. While there is no way to confirm a model structure definitively, unsuccessful models can be eliminated from consideration by their inability to fit the available data for any set of parameters. Since models and parameters are tested simultaneously, the MCMC method for diagnosing parameter non-identifiability may be also be useful for model selection (Siekmann et al., 2012). When the available data lacks the power to constrain the parameters of a model, it is likely that many other models of comparable complexity will also easily fit that data. Therefore, diagnosing identifiability comes as a first step in the model selection process whereby potential models are discarded from consideration if they cannot be constrained by the data. Detecting when model parameters are not identifiable can indicate situations in which model selection is also compromised.

2.3.2 Relationship to previous work

The strong inter-relationships between experimental design, model selection, and parameter estimation have been rediscovered in many fields, including econometrics (Koopmans, 1949; Rothenberg, 1971), process industries (Gustavsson, 1975;Chappell and Godfrey, 1992), systems and control engineering (Eykhoff, 1964; Lee, R.C.K., 1964), and, more recently, systems biology

(Audoly et al., 2001; Chis et al., 2011). These ideas have been rigorously systematized into a unified discipline called system identification (Eykhoff, 1974; Ljung, 1987; Goodwin and Payne, 1977; Pronzato and Walter, 1997). The parameter identifiability aspect of system identification has been explored extensively in the control theory literature (Astrom and Bellman 1970; Grewal and Glover, 1974; Cobelli and DiStefano III, 1980). Recently, there has been a surge of interest in questions of parameter identifiability for models of large biological systems, such as genetic, metabolic, biochemical, and ecological networks, and signal transduction, cell cycle, and pharmacodynamic pathways (Audoly et al., 2001; Cheung et al., 2013; Chis et al., 2011; Hengle et al., 2007). For these complex, interconnected systems, the parameter compensations that result in non-identifiability are possible because of the large number of parameters required to model them.

There is a large literature on fitting binding curves of single- and multisite receptors using models such as the Hill model and the Adair model (Hill, 1913; Adair, 1925; Klotz, 1997; Wyman and Gill, 1990; Forsen and Linse, 1999). However, there has been relatively little treatment of identifiability for simple biochemical systems (Hines, 2013; Raue et al., 2009; Johnson et al, 2009a; Bruno et al., 2005; Kienker, 1989). We show here that the parameters of even extremely simple models are often not identifiable, suggesting that this problem may be more widespread than is generally appreciated. Parameter non-identifiability in biochemical systems was investigated by J.G. Reich and colleagues in the 1970s (Reich, 1974, Reich et al., 1974a, Reich et al., 1974b, Reich and Zinke, 1974). In this work, the authors address difficulties when analyzing ligand binding data as well as kinetic data. They proposed methods of calculating parameter sensitivity to detect redundant parameters (non-identifiability) and use these methods to compare various binding models (such as those shown in Figure 1) to quantify the information content in a binding curve. Their work predates modern computing power and the widespread use of efficient sampling algorithms such as MCMC.

Although we have focused on curve-fitting applications, the MCMC method can be applied to any model for which posterior probabilities can be calculated. For example, stochastic process models are commonly used for modeling the dynamics of molecules. Markov Models and Hidden Markov Models have been used to understand the conformational dynamics of ion channels (Qin et al., 1997), molecular motors (Mullner et al., 2010), and ligand-binding proteins (Stigler and Rief, 2012). In these settings, the stochastic properties of single molecule time series are used to constrain model parameters (transition rates between states). The model parameters are estimated by maximizing the likelihood of the data (or the posterior probability). Commonly, a point estimate of the parameters of a candidate model is calculated (Rabiner, 1989), but such an approach does not indicate whether these parameters are uniquely constrained by a particular time series. In contrast, MCMC samples the full posterior distributions and thus provides an indication of non-identifiability. This approach has been applied to the study of ion channel gating (Siekmann et al., 2012) and may become a powerful method for developing useful models of molecular processes.

2.3.3 Computational advantages of MCMC

The Bayesian framework presented here has clear advantages over alternative methods of diagnosing parameter identifiability. One approach might be to examine the sensitivity of the fit to changes in the parameters, using a variety of matrix based methods. We showed that this approach can only be applied in special cases and can even misleadingly suggest reasonable parameter estimates in the presence of realistic experimental noise. It is necessary to directly explore the full range of the parameter space that leads to good agreement with the data. Therefore, an alternative approach might be to directly compute the error between the data and model for an entire parameter space. This approach works well for simple problems, but is not feasible for large models. For a K-dimensional model, computation of N points for each parameter requires $O(N^K)$ error calculations (here the notation $O(f(N, K))$ specifies that as a function of N and K, the number of computations is on the order of $f(N, K)$). Obviously this exponential explosion makes larger models impractical. An alternative is to consider just the pair-wise parameter correlations for all model parameters and compute the total error. This approach was taken in Figs. 2.2C and 2.3C and has been employed previously (Johnson et al., 2009b, Johnson et al., 2009a). This method is limited, since errors are calculated on a large joint-error-surface, and therefore computational effort is wasted in regions of parameter space that yield poor fits to the data. In addition each total-error computation involves finding the best-fit point of the other parameters and thus entails $O(N^2K^2)$ repetitions of some optimization algorithm, which itself might involve many iterations to reach convergence. In contrast, MCMC focuses computational effort in the region of parameter space that is relevant to the data. Further, posterior estimation requires only $O(NK)$ repetitions of a simple calculation of posterior probability. Such Bayesian methods have recently been embraced by the systems biology community, where inference is routinely conducted on models containing more than 70 free parameters (Eydgahi et al., 2013; Klinke, 2009; Battogtokh et al., 2002). Using MCMC to sample posterior distributions yields not only accurate parameter estimates in high dimensional spaces, but also provides information regarding identifiability and nonlinear parameter correlations. Our MCMC implementations use the Metropolis-Hastings algorithm, which is conceptually simple, but is not optimal for high-dimensional problems. Fortunately, more sophisticated MCMC algorithms have been developed (Neal, 2010; Girolami and Calderhead, 2011).

We briefly mention some of the practical considerations that must be noted when using MCMC to estimate posterior distributions. Figure 2.4C shows the parameter trajectories of MCMC samples from an identifiable model. Each of the parameters are initialized at an arbitrary value and these trajectories visualize how parameter estimates move toward regions of high posterior probability. Once these large movements of the parameters cease, the chain makes transitions only in proportion to the posterior probability and the Markov chain is said to have reached stationarity (or converged). After convergence, all subsequent transitions of the chain produce iid samples from the posterior and can be used for parameter estimation. The iterations preceding convergence are termed the burn-in period and these values are discarded. The MCMC samples visualized in Figs. 2.2D, 2.3D, the histograms of Figure 2.4, and Figure 2.5 have excluded the burn-in samples. It is important to determine when the Markov chain has reached stationarity and many methods can be used. Most simply, one could assess convergence by visual inspection: the trajectories in Figure 2.4C seem to have converged by 200 iterations. More rigorous methods are desirable and many have been developed; we point the reader to (Gelman and Rubin, 1992; Geweke, 1992). We also direct the reader to (Gilks et al., 1996) for a discussion of chain mixing efficiency and the effect on burn-in time. For non-specialists and those interested in implementing MCMC sampling, there are two excellent introductory handbooks (Brooks et al., 2011; Gilks et al., 1996b) that provide practical advice and guidance, and include numerous case-studies of MCMC applied in diverse fields such as epidemiology, genetics, archaeology, ecology, and image analysis.

2.3.4 Parameter identifiability and experimental design

The tools described here for diagnosing parameter identifiability can be a useful component of the experimental design process. Figures 2.4 and 2.6 present potential signals that might be used to constrain different kinetic schemes. While previous work has addressed model discrimination with macroscopic kinetic time series (Celentano and Hawkes, 2005), a Bayesian approach provides a direct assessment of identifiability. By sampling the posterior distribution using MCMC, we showed that the model with four parameters (Figure 2.4) is uniquely constrained by the data. Conversely, the model with eight parameters (Figure 2.6) is non-identifiable when constrained by the data, and inferences about the properties of this model would be meaningless. In the latter case, we may reject the initial model in favor of one with fewer parameters, although the parameters of the simpler model may lack the required mechanistic significance. Alternatively, if this model is motivated by specific phenomenological considerations, then we may be resistant to reject it. To derive meaningful mechanistic conclusions from this system we must then redirect our experimental efforts in one of three ways: i) by performing the same experiment, but collecting data with sufficiently higher signal-to-noise ratio (in the case of practical non-identifiability), ii) by collecting other types of data using existing approaches, or iii) by devising novel experiments that generate observable signals with greater constraining power.

Many examples exist to illustrate the power that new types of data bring to our ability to quantitatively model biophysical phenomena. For example, it is particularly difficult to differentiate binding events from conformational changes in ligand-gated ion channels when only macroscopic ionic current measurements are available (Colquhoun, 1998). Recently, the Benndorf group has pioneered an approach in which channel opening is measured electrophysiologically, while ligand-binding events are detected simultaneously by fluorescence methods (Kusch et al., 2011). A second example is the role that single channel recording has had on the development of mechanistic models of ion channel gating. For example, we demonstrated that the parameters of the model in Figure 2.6 are non-identifiable when constrained by macroscopic current relaxations. However, single channel recordings can be used to constrain models of this complexity (Colquhoun and Sakmann, 1985; Sakmann and Neher, 1995). Another example is the transformative role of gating current measurements in elucidating mechanisms in voltage-gated ion channels (Armstrong and Bezanilla, 1973; Keynes and Rojas, 1974). Complementing ionic current measurements with gating currents can reveal parts of the state space of a model that are difficult to distinguish with ionic currents alone. The constraining power of the additional data reduces compensation in the model parameters and results in an identifiable model. Returning to the problem of calcium binding to CaM, we showed that the parameters of the four-site sequential model (Figure 2.1A) are not identifiable when constrained by measurements of the net occupancy of CaMs four metal binding sites. Our analysis indicates that enormous parameter uncertainties will result from fitting typical binding data, even data with noise that is lower than is practically achievable. However, this barrier may be overcome by performing experiments that quantify the site-specific occupancy of the four binding sites in CaM as a function of calcium concentration (diCera, 1995). Our results indicate that, given the frequent occurrence of non-unique parameter estimation, analyzing parameter identifiability should become a standard component of the experimental design process.

Chapter 3

Inferring Subunit Stoichiometry From Single Molecule Photobleaching

The majority of the text and figures presented here have been published previously in the *Journal of General Physiology*:

Hines, K.E. (2013). Inferring Subunit Stoichiometry From Single Molecule Photobleaching. Journal of General Physiology. 141(6):737-46.

Abstract Single molecule photobleaching is a powerful tool for determining the stoichiometry of protein complexes. By attaching fluorophores to proteins of interest, the number of associated subunits in a complex can be deduced by imaging single molecules and counting fluorophore photobleaching steps. Because some bleaching steps might be unobserved, the ensemble of steps will be binomially distributed. In this work, it is shown that inferring the true composition of a complex from such data is nontrivial because binomially distributed observations present an ill-posed inference problem. That is, a unique and optimal estimate of the relevant parameters cannot be extracted from the observations. Because of this, a method has not been firmly established to quantify confidence when using this technique. This paper presents a general inference model for interpreting such data and provides methods for

accurately estimating parameter confidence. The formalization and methods presented here provide a rigorous analytical basis for this pervasive experimental tool.

3.1 Introduction

The method of single molecule photobleaching has become a popular tool to examine stoichiometry and oligimerization of protein complexes. In recent work, this method has been used to determine the stoichiometry of a great variety of transmembrane proteins such as ligand-gated ion channels (Reiner et al., 2012; Ulbrich and Isacoff, 2008; Yu et al. 2012), voltage-gated ion channels (Nakajo et al., 2010), mechano-sensitive channels (Coste et al., 2012) and calcium-release-activated calcium channels (Demuro et al., 2011; Ji et al., 2008). Additionally, this method has been used to examine complexes of other types of proteins such as β-Amyloid (Ding et al., 2009), helicase loader protein (Arumugam et al., 2009), and toxin Cry1Aa (Groulx et al., 2011), among many others. The approach consists of attaching a fluorescent probe (typically GFP or a variant) to a protein subunit of interest and imaging single molecules. After sufficient excitation, a fluorophore will bleach, resulting in a step-wise decrease in observed fluorescence. Then, by simply counting the number of these bleaching steps, one can observe how many fluorophores were imaged and thus how many subunits, n, were associated in the observed complex. However, there is a non-zero probability, 1 - θ, that any given flurophore will already be bleached (or otherwise unobserved) and thus less than the highest possible number of fluorescence decreases will be observed. Stated differently, the parameter θ is the probability of successfully observing each possible photobleaching event. As noted by the originators of this method, the resulting observations are drawn from a binomial distribution (Ulbrich and Isacoff, 2007), and thus the highest observed number of bleaching steps is the *minimum* number of subunits in the complex.

As an example, consider the data shown in Figure 3.1 A and B. Here, I have reproduced the distributions of observed bleaching steps reported in (Ulbrich and Isacoff, 2007) and (Coste et al., 2012), respectively. In both of these studies, the investigators are using the method of single molecule photobleaching to quantify the assembly of alpha subunits of the cyclic nucleotide-gated ion channel (CNG). These experiments are performed on the same protein, and both show that the highest observed number of bleaching steps is four. Note that these distributions are quite different, as preparation variability between the two experimental groups has likely led to differences in fluorophore pre-bleaching (ie., differences in θ). In (Ulbrich and Isacoff, 2007), the authors report that $\theta = .8$ and in (Coste et al., 2012) the value is not reported, but I estimate it to be approximately .5, which is not much lower than other reported values, such as .53 (McGuire et al., 2012). It is unclear how the differences in these distributions (and in θ) should impact the interpretation of these results. Both of these distributions provide evidence that the CNG channel is a tetramer, but to what extent does one of these distributions provide *better* evidence in support of this conclusion?

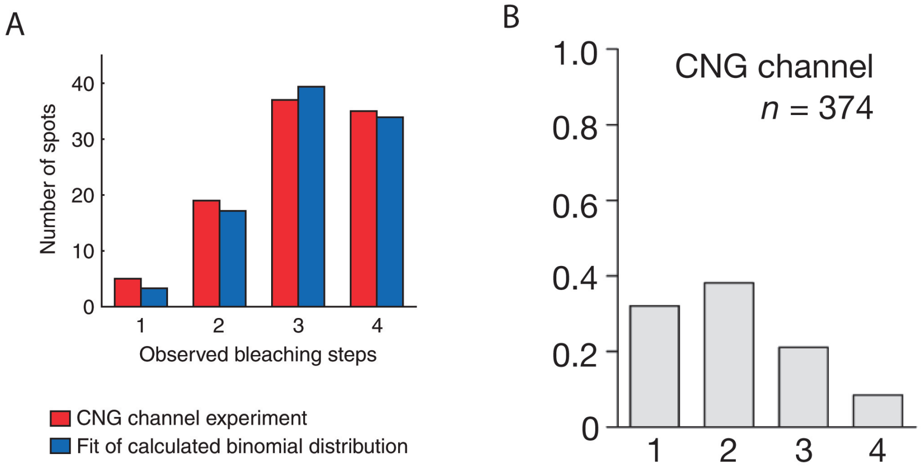

Figure 3.1: Example distributions reproduced from the literature. The observed distributions reported in (Ulbrich and Isacoff, 2007) A and (Coste et al., 2012) B, when using the method of single molecule photobleaching to assess the stoichiometry of the cyclic nucleotide-gated ion channel. In both of these distributions, the highest observed number of bleaching events is four. However, note that these distributions are quite different, likely due to preparation variability. A method has not been established which takes into account the properties of the observations in order to accurately accept and reject hypotheses.

It is not immediately obvious how to determine the confidence with which the number of subunits can be inferred from these observations. In particular, it is possible that the true n is actually larger than the highest observed number of bleaching steps, but due to the finite sample size, the true tails of the distribution were not observed. Alternatively, the data collection algorithm might have resulted in artifactual observations, causing an overestimation of n. A method has not been firmly established to determine whether parameter estimates are unique and the confidence with which parameters can be inferred from this data. I show that this inference is non-trivial because binomial distributions present an ill-posed inference problem: there does not exist a unique combination of n and θ which could have produced a particular set of observations. As a result, it may be highly likely that this data is misinterpreted. To resolve this disparity, I present a generalized method of inference which provides accurate estimates of parameter confidence. The methods developed here will prevent misinterpretation and will yield more fruitful experimentation and accurate conclusions.

3.2 Results and Discussion

3.2.1 Bayesian Inference

Since the analysis presented in this paper employs Bayesian inference, this section provides a brief tutorial. Suppose that we have some probability model with m parameters $\{\theta_1, \theta_2, ..., \theta_m\} = \vec{\theta}$. This model will be denoted $p(y_i|\vec{\theta})$ for any observable y_i and quantifies the probability of observing some y_i given the values of parameters $\vec{\theta}$. If we gather observations $\{y_1, y_2, ..., y_N\}$, denoted y_N, then the aim of statistical inference is to use observations y_N to infer the true values of parameters $\vec{\theta}$. While it may be simple to obtain a single, optimal estimate of the parameters given the data, the goal of Bayesian inference is to consider all possible values of the parameters and quantify which regions of parameter space are most consistent with the observations. This is achieved by constructing a probability distribution over the parameter space (the posterior distribution), where areas of higher posterior probability are in better agreement with the data than areas of lower posterior probability. In this way, our uncertainty in estimating the parameters is captured by the posterior distribution of the parameters given the data, $p(\vec{\theta}|y_N)$. Posterior distributions can be calculated from $p(y_i|\vec{\theta})$, the likehood of observing y_i given $\vec{\theta}$, and $p(\vec{\theta})$, the prior distribution of the parameters. Using the posterior distribution, we are able to quantify the full uncertainty in all model parameters.

3.2.2 Binomial Distributions and Ill-Posed Inference

If k fluorescently labelled protein subunits are associated together, then one might expect to observe k photobleaching steps. However, each fluorophore may already be bleached, with probability $1 - \theta$. The likelihood of observing k bleaching steps, if a total of n steps are possible, will follow a binomial distribution :

$$\begin{aligned} p(k) &= p(k|n, \theta) = \mathrm{Bn(n,}\theta) && (3.1)\\ &= \frac{n!}{(n-k)!k!}\theta^k(1-\theta)^{n-k}. && (3.2) \end{aligned}$$

Consider that we have one observed number of bleaching steps, y_i, and wish to estimate θ and n. Further, we wish to estimate the full distributions over parameters θ and n that are most consistent with this observation. From Bayes' rule, we calculate this posterior probability distribution as

$$\begin{aligned} p(\theta, n|y_i) &\propto p(y_i|\theta, n)p(\theta)p(n) && (3.3)\\ &\propto \frac{n!}{(n-y_i)!y_i!}\theta^{y_i}(1-\theta)^{(n-y_i)}p(\theta)p(n). && (3.4) \end{aligned}$$

where $p(\theta)$ and p(n) are the prior distributions over the values taken by parameters θ and n. If N observations are independent, this proceeds similarly for the full set y_N:

$$p(\theta, n|y_N) \propto \tag{3.5}$$

$$\prod_{i=1}^{N} \frac{n!}{(n-y_i)!y_i!} \theta^{y_i}(1-\theta)^{(n-y_i)} p(\theta)p(n). \tag{3.6}$$

As an example of posterior inference, imagine that we have observations drawn from a binomial distribution with a known n and we wish to estimate θ. Since we suppose that n is known, the joint posterior distribution (equation 3.6) reduces to just the posterior distribution of θ:

$$p(\theta|y_N, n) = \tag{3.7}$$

$$\prod_{i=1}^{N} \frac{n!}{(n-y_i)!y_i!} \theta^{y_i}(1-\theta)^{(n-y_i)} p(\theta). \tag{3.8}$$

We need to decide upon a form for the prior distribution $p(\theta)$. Since θ is the probability of a binary event, a useful and flexible form for $p(\theta)$ will be the Beta distribution, Be(a,b). This distribution is defined on the interval [0,1] and has two parameters, a and b. If we have little prior information about θ, then setting $a = b = 1$ results in a flat prior distribution. If, however, we have a strong guess about θ, then parameters a and b can be chosen to properly reflect our prior belief. In either case, the posterior distribution is

$$p(\theta|y_N, n) \propto \tag{3.9}$$

$$\prod_{i=1}^{N} \frac{n!}{(n-y_i)!y_i!}\theta^{y_i}(1-\theta)^{(n-y_i)}p(\theta) = \tag{3.10}$$

$$\prod_{i=1}^{N} \frac{n!}{(n-y_i)!y_i!}\theta^{y_i}(1-\theta)^{(n-y_i)}\frac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)}, \tag{3.11}$$

where $\beta(a,b)$ is the proper normalization constant. The form of this posterior simplifies to a useful result:

$$p(\theta|n, y_N) \propto \tag{3.12}$$

$$\prod_{i=1}^{N} \frac{n!}{(n-y_i)!y_i!}\theta^{y_i}(1-\theta)^{(n-y_i)}\frac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(\text{a,b})} \tag{3.13}$$

$$= \prod_{i=1}^{N} \frac{1}{\beta(\text{a,b})}\frac{n!}{(n-y_i)!y_i!}\theta^{y_i+a-1}(1-\theta)^{n-y_i+b-1} \tag{3.14}$$

$$\propto \prod_{i=1}^{N} \theta^{y_i+a-1}(1-\theta)^{n-y_i+b-1} \tag{3.15}$$

$$= \theta^{\sum_{i=1}^{N}(y_i+a-1)}(1-\theta)^{\sum_{i=1}^{N}(n-y_i+b-1)} \tag{3.16}$$

It can be seen that this posterior of θ is also a Beta distribution,

$$p(\theta|n, y_N) \propto \text{Be(A,B)} \tag{3.17}$$

$$\text{where A } = \sum_{i=1}^{N} y_i + a - 1 \tag{3.18}$$

$$\text{and B} = \sum_{i=1}^{N} n - y_i + b - 1. \tag{3.19}$$

Therefore, if data are drawn from a binomial distribution, the posterior distribution of θ (with respect to a fixed n) will be a Beta distribution with parameters $A = \sum_{i=1}^{N} y_i + a - 1$ and $B = \sum_{i=1}^{N} n - y_i + b - 1$. Figure 3.2 is an example of the posterior distribution of θ for some simulated data. The black vertical line is simply the estimate of θ that one would calculate by varying the value of θ to find a best fit to the model Bn(4,θ): this is the maximum likelihood estimate (MLE). The other curves in Figure 3.2 are the posterior probabilities of θ for two hypothetical datasets of different sizes. Note that in the absence of strong prior information, the maximum value of the posterior distribution (the *maximum a posteriori* (MAP) estimate) will equal the value of θ that we estimate by finding the best fit to the data (the MLE). In this way, the full posterior distribution over the parameter not only provides an optimal point estimate (MAP), but also provides a confidence about the full range of the parameter and which values are consistent with the data. As we would expect, as the number of observations increases, the resulting posterior distribution will become more narrow and we will have less uncertainty regarding the true value of θ. Finally, note that the estimates of θ will depend on the value of n, and that the conditional posterior distribution, $p(\theta|y_N, n)$, defines a family of distributions for various values of n. This result will be useful later.

For the experimental setting of single molecule photobleaching, n is not known, but instead needs to inferred from the data. After gathering some observations, y_N, we can determine the highest observed number of bleaching

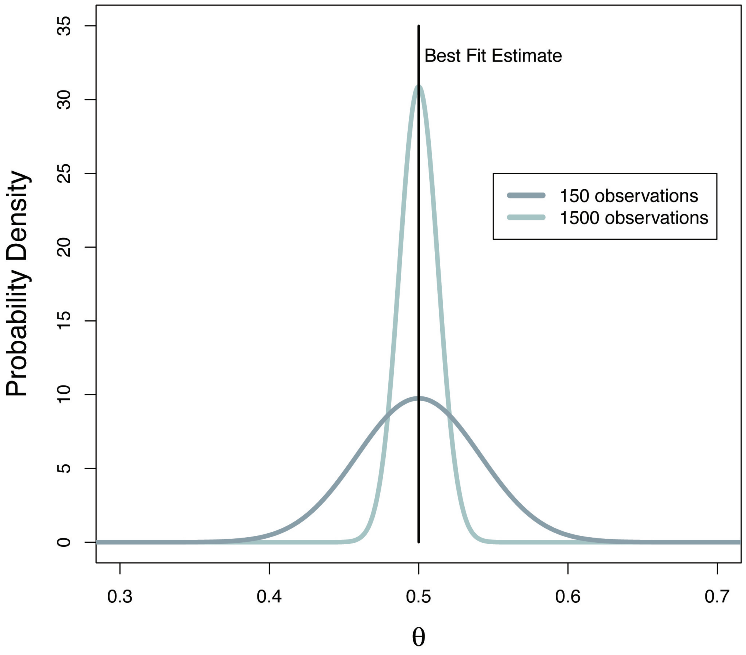

Figure 3.2: Posterior probability distribution of θ. An example of the posterior probability of θ for hypothetical datasets. The vertical black line represents the optimal point estimate of θ that one would calculate by curve-fitting (the MLE). The other curves are the posterior probability distributions of θ for different amounts of data. Note that as the number of observations increase, the posterior distribution is narrowed as our confidence about the true value is improved. Also note that the maximum value of posterior probability coincides with the MLE that we would calculate by curve fitting. Calculating the posterior distribution over parameters provides not only an optimal point estimate, but also a quantification of parameter uncertainty.

steps, $\hat{k}$, and be tempted to conclude that $n = \hat{k}$. Before doing this, we will want a way to establish that $n = \hat{k}$ is highly supported by the data and that all other $n > \hat{k}$ are not supported by the data. We want to calculate $p(n|y_N)$, the marginal posterior distribution over n. This is the probability (over all n) of a particular n having given rise to the observations. We can directly calculate the marginal distribution of n for this model. Consider a single observation $y_i = k$. The joint posterior is

$$p(\theta, n|k) = \frac{n!}{(n-k)!k!}\theta^k(1-\theta)^{n-k}p(\theta)p(n) \tag{3.20}$$

Since θ represents the probability of a binary event, we use a Beta distribution as the prior, $p(\theta)$=Beta(a,b), and set the prior on n as a bounded uniform distribution. As mentioned previously, we never know θ with certainty, so we must consider all possible values of θ for each n. The marginal posterior of n is then found by integrating over θ,

$$p(n|k) = \int_0^1 p(\theta, n|k)p(\theta)p(n)d\theta \tag{3.21}$$

$$\propto \int_0^1 \frac{n!}{(n-k)!k!}\theta^k(1-\theta)^{n-k}\theta^a(1-\theta)^b d\theta \tag{3.22}$$

$$= \frac{(n-1)!}{(n-k)!k!}\frac{\Gamma(k+a)\Gamma(n-k+b)}{\Gamma(n+a+b)}, \text{for } n \geq k \tag{3.23}$$

where $\Gamma()$ is the gamma function. The marginal posterior of n takes the form of this ratio of gamma functions and it can be seen that this function is zero for $n < k$, is maximized at k, and is monotonically decreasing for $n > k$. For many independent observations, the relevant posterior, $p(n|y_N)$, is just a product of these functions and will have the general property of being maximized at the largest observed y_i and rapidly decrease for $n > \hat{k}$.

Note that the marginal distribution of n (equation 3.23) will depend only on the largest observed y_i. Consider the case that the true n is larger than $\hat{k}$, but due to the finite sample size, no evidence of the true n was observed. In this case, the posterior distribution will always be peaked at the smallest n which can explain the data, and any greater n will have much smaller posterior probability. This provides little ability to compare the evidence from, say, Figure 3.1 A and B. We can ignore the Bayesian approach used thus far and simply calculate the maximum likelihood estimate for n given k and again see that the likelihood is always maximized at the smallest n that can account for the data. Therefore, typical methods of estimation will fail in this pursuit, and it is worth understanding why this is the case. This undesirable property stems from the fact that this inference problem is *ill-posed*: there is generally not a unique solution for n and θ for a given dataset. To visualize this, we can compute the joint posterior distribution (equation 3.20) for a simulated data. This joint posterior is plotted in Figure 3.3A for a region of the parameter space in θ and n and areas of lighter color correspond to areas of higher posterior probability (analogous to lower error between the data and the model). For example, if we examine $p(\theta|n = 4, y_N)$ then a horizontal slice through the joint posterior (at n=4) corresponds to our estimate of θ given that $n = 4$ and this distribution is peaked around .6. Notice that for each $n > 4$, the estimate of θ systematically shifts downward to lower values. This must be the case, since if a binomial process of $n = 10$ somehow generated the data in Figure 3.1, then the failure probability, $1 - \theta$, would have to be quite high to have generated no observations exceeding $y_i = 4$. As a consequence, notice that the joint posterior (Figure 3.2A) is highly structured, and it is possible for *any* arbitrary n to have generated the data with a compensatory decrease in θ. Further, the *most* probable estimate for n will always be the smallest possible one, regardless of the observed distribution. Due to this, methods which rely solely on likelihood calculation will not be able to discern the most accurate estimate of these parameters.

To demonstrate how this ill-posed inference impairs our ability to learn n from data, in Figure 3.3B and C I have simulated data meant to mimic the range seen in Figure 3.1A and B by drawing from a binomial distribution with $n = 4$ and θ equal to .8 (B) and .5 (C). In each case we are tempted to conclude the true n is four, but can we make this assertion with equal vigor in both instances? An obvious approach is fitting binomial distributions to the data and assessing the quality of fit. The circles in Figure 3.3 represent the best fit to a binomial distribution with $n = 4$ and it is clear that these fit the data well in both cases and that we are able to accurately estimate the optimal value of θ. However, in order to be confident about the assertion that $n = 4$, we must ask whether these fits are unique. The crosses in Figure 3.3 show the best fit to a binomial process with $n = 5$. In B, it is immediately obvious that even the best fit is a poor match to the data: the n=5 binomial distribution underestimates the number of observed 3- and 4-bleaching steps and also predicts that roughly 10% of the data should have reflected 5-bleaching steps, whereas no 5-bleaching steps are observed. In this case, it is very obvious that $n = 4$. In C, we cannot be so certain. While the $n = 4$ binomial distribution certainly provides a good fit to the data, the $n = 5$ model also fits the data quite well for all observed bleaching steps. Further, the $n = 5$ fit predicts that only 1% of the data should reflect 5-bleaching steps, and thus we might not have seen any simply due to the finite sample size. In this case, fits to the data are not unique and n and θ can compensate to produce identically good fits. This stems directly from the fact that this inference problem is ill-posed, as depicted in the joint posterior distribution (Figure 3.3A). However, note that the possibility of this underestimation of n depends very strongly on the value of θ and qualitatively we can be more confident in the data in B than C. The methods proposed in the next section quantify this confidence.

3.2.3 Parameter Confidence

Returning to example data, such as that in Figure 3.3 B or C, suppose we have observed some maximum number of bleaching steps, $\hat{k}$, and are tempted to conclude that $n = \hat{k}$, but want to consider the irksome possibility that $n > \hat{k}$, though we did not observe any evidence of it. We would like to make a statement to the effect of : Given N observations less than or equal to $\hat{k}$, we can conclude with confidence α that the true n is less than $\hat{k} + 1$. The

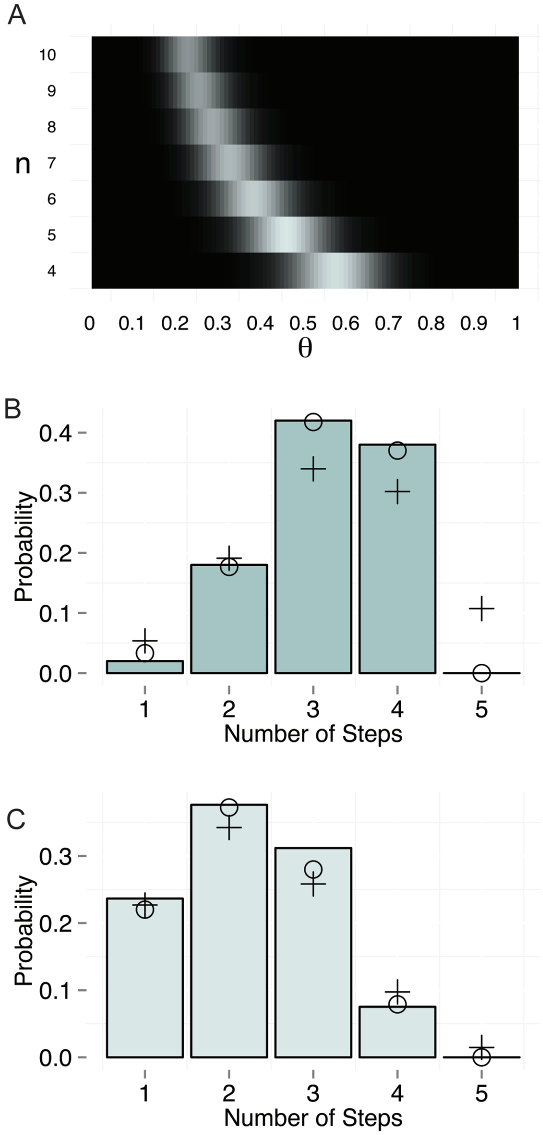

Figure 3.3: Ill-posed inference. (A) Joint posterior distribution of n and θ, given simulated data, y_N. Areas of lighter color reflect areas of higher posterior probability. (B) Data drawn from a binomial distribution with $n = 4$ and $\theta = .8$. (C) Data drawn from a binomial distribution with $n = 4$ and $\theta = .5$. In B and C, circles (o) represent the best fit to a binomial distribution with $n = 4$ and crosses (+) represent the best fit to a binomial distribution with $n = 5$. In C, the best fit to the data for $n = 4$ and $n = 5$ are equally good because such fits are not unique.

strategy I propose is similar, in spirit, to classical hypothesis testing, where the null hypothesis is that $n > \hat{k}$ and 1-α quantifies the probability of observing $\hat{k}$ under the null hypothesis.

As the null hypothesis, assume that $n = \hat{k} + 1$, but we simply did not observe any $y_i = \hat{k} + 1$ due to finite sample size. For simplicity, assume (unrealistically) that we have an exact point estimate of θ for $n = \hat{k} + 1$, denoted $\hat{\theta}$ (this assumption will be relaxed later). Then the probability of observing an event of size $\hat{k} + 1$ is

$$p(y_i = (\hat{k}+1)|\hat{k}+1, \hat{\theta}) = \frac{(\hat{k}+1)!}{((\hat{k}+1)-(\hat{k}+1))!(\hat{k}+1)!}\hat{\theta}^{(\hat{k}+1)}(1-\hat{\theta})^{((\hat{k}+1)-(\hat{k}+1))} \tag{3.24}$$

$$= \frac{(\hat{k}+1)!}{(\hat{k}+1)!}\hat{\theta}^{(\hat{k}+1)}(1-\hat{\theta})^{0} \tag{3.25}$$

$$= \hat{\theta}^{\hat{k}+1}. \tag{3.26}$$

We then need to calculate the probability of not seeing this event, given that we have N observations. To do this, we consider the sampling distribution of events $y_i = \hat{k} + 1$ under the null hypothesis, Bn($\hat{k}$ + 1,$\hat{\theta}$). This results in another binomial distribution, Bn(N,$\hat{\theta}^{\hat{k}+1}$), where there are N chances of observing the event and the probability of the event is $\hat{\theta}^{\hat{k}+1}$. Then the probability of $\hat{k}$ being the highest observed y_i is $p(0|N, \hat{\theta}^{\hat{k}+1})$ and our estimate of confidence, α, is $1 - p(0|N, \hat{\theta}^{\hat{k}+1})$:

$$\alpha = 1 - p(0|N, \hat{\theta}^{\hat{k}+1}) \tag{3.27}$$

$$= 1 - \frac{N!}{(N-0)!0!}(\hat{\theta}^{\hat{k}+1})^0(1 - (\hat{\theta}^{\hat{k}+1}))^N \tag{3.28}$$

$$= 1 - (1 - \hat{\theta}^{\hat{k}+1})^N. \tag{3.29}$$

As an approximate guide for experimental design, we can systematically explore the space of θ in order to understand how probable this underestimation actually is. In Figure 3.4A, I have plotted α (confidence) for a region of θ and N and for a fixed value of $\hat{k}$=4. Again, smaller values of α mean that there is a higher probability of not observing the true tails of distribution under the null hypothesis. For smaller values of α, we cannot be confident that a dataset with a similar θ and N was not drawn from a binomial distribution which was larger than indicated by the data. For visual ease, the colormap in Figure 3.4A focuses on several contours of α and the colors threshold all α values to where they lie within these regimes. From this systematic exploration, some useful insights emerge. As we might have guessed, for large θ, the probability of underestimating the true n is trivially small, even for small datasets. However, for θ in the range of only .5, which has been seen multiple times in the literature, this possibility is not so rare. For concreteness, Figure 3.4B shows α as a function of sample size for two values of θ. For high θ, we can have high confidence in a conclusion even for a dataset of size 25. Conversely, if θ is .5, then a dataset of the same size would lead to the wrong conclusion with probability approximately $\frac{1}{2}$. Returning to the data from Figure 3.1, we can now (approximately) assess the strength of each of these data sources. We can be very confident in these data sources as $1-\alpha < 10^{-6}$ in both instances. Fortunately, these two examples from the literature both provide reliable evidence that the CNG channel is a tetramer, though without using such methods, we would have been unable to quantify this confidence.

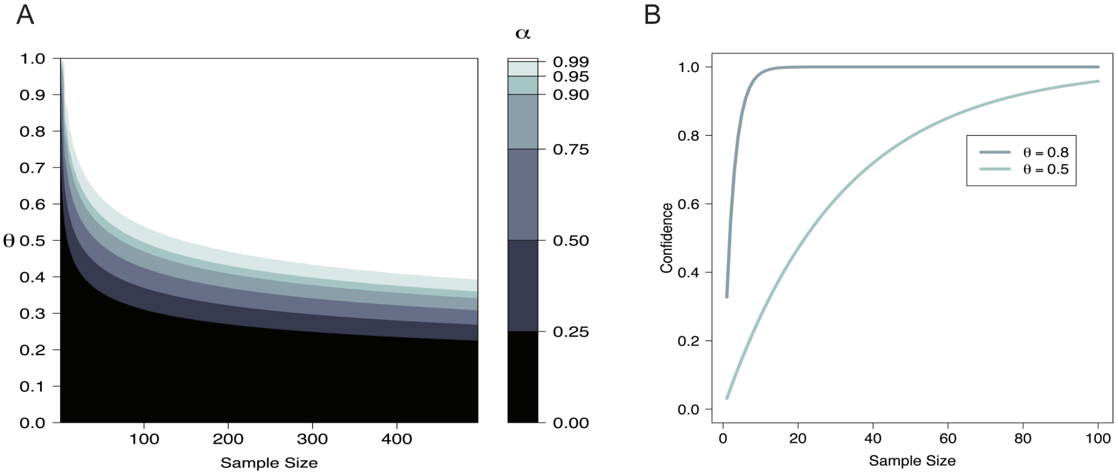

Figure 3.4: Estimated parameter confidence. (A) Estimated α for various θ and sample sizes. Value of α is represented by the colormap. For simplicity, contours of α are shown and the color of each region indicates areas where α lies between these contours. (B) α as a function of sample size for two values of θ.

It is also important to establish that our estimate of confidence with respect to the hypothesis $n = \hat{k} + 1$ is a lower bound on all conceivable hypotheses $n > \hat{k} + 1$. For simplicity, we will first consider the potential null hypothesis $n = \hat{k} + 2$. Again, we are assuming that we have an optimal estimate $\hat{\theta}$, but now with respect to the hypothesis $\mathrm{Bn}(\hat{k} + 2, \hat{\theta})$. As above, the probability of observing an event $y_i = \hat{k} + 2$ is $\hat{\theta}^{\hat{k}+2}$. Given that we have N observations, the probability of observing zero events of size $y_i = \hat{k} + 2$ is,

$$p(0|N, \hat{\theta}^{\hat{k}+2}) = (1 - \hat{\theta}^{\hat{k}+2})^N. \tag{3.30}$$

The probability of observing an event of size $y_i = \hat{k} + 1$,

$$\begin{aligned} p(\hat{k} + 1|\hat{k} + 2, \hat{\theta}) &= \frac{(\hat{k} + 2)!}{((\hat{k} + 2) - (\hat{k} + 1))!(\hat{k} + 2)!} \hat{\theta}^{\hat{k}+1}(1 - \hat{\theta})^{(\hat{k}+2)-(\hat{k}+1)} && (3.31) \\ &= (\hat{k} + 2)\hat{\theta}^{\hat{k}+1}(1 - \hat{\theta}). && (3.32) \end{aligned}$$

The probability of observing exactly zero of these events, given a total N observations is,

$$p(0|N, (\hat{k} + 2)\hat{\theta}^{\hat{k}+2}(1 - \hat{\theta})) = (1 - (\hat{k} + 2)\hat{\theta}^{\hat{k}+1}(1 - \hat{\theta}))^N. \tag{3.33}$$

The probabilty of seeing no observations of size $\hat{k}$+1 or $\hat{k}$+2 is just the product of equations 3.30 and 3.33. Therefore, our confidence that true n is not $\hat{k} + 2$ goes as,

$$\alpha = 1 - (1 - \hat{\theta}^{\hat{k}+2})^N (1 - (\hat{k}+2)\hat{\theta}^{\hat{k}+1}(1-\hat{\theta}))^N. \tag{3.34}$$

Generally, the estimate $\hat{\theta}$ used in equation 3.30 will be less than that of equation 3.29 (see Figure 3.3A). However, the confidence estimate in equation 3.34 involves multiplication with an additional term than equation 3.29. Therefore, the confidence estimated when considering the hypothesis $n = \hat{k}+2$ will always be higher than that for the hypothesis $n = \hat{k}+1$. This is visualized in Figure 3.5 where confidence is plotted as a function of sample size for the null hypothesis $n = \hat{k}+1$ in black and for the null hypothesis $n = \hat{k}+2$ in teal. Clearly, the estimate of confidence with respect to $\hat{k}+1$ is the most conservative estimate. It is easy to see that this relationship will persist for all $n > \hat{k}+1$. Due to this, we only need to calculate confidence with respect to $\hat{k}+1$, as this provides a lower bound on confidence with respect to all possible $n > \hat{k}$.

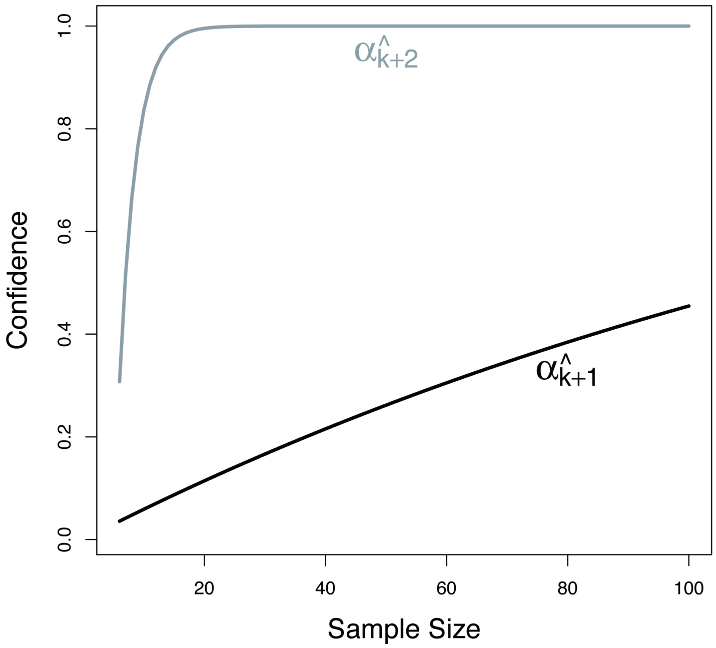

Figure 3.5: Comparison of parameter confidence when considering multiple models. The black curve is a plot of confidence versus sample size for the null hypothesis $n = \hat{k} + 1$. The teal curve is the parameter confidence for the null hypothesis $n = \hat{k} + 2$. It is clear that only the hypothesis $n = \hat{k} + 1$ needs to be considered and will result an estimate of confidence which is a lower bound on all possible hypotheses.

The previous discussion provides a notion of confidence only if we know the value of θ exactly. As this is never the case (Figure 3.2), we need to generalize equation 3.29 to include our uncertainty in the value of θ. This uncertainty is quantified by the conditional posterior distribution, $p(\theta|y_N, \hat{k}+1)$, with respect to the null hypothesis $n = \hat{k}+1$. Our estimate of alpha should consider all possible values of θ, weighted by their posterior probability. In particular,

$$\begin{aligned}
\alpha =& 1 - \int_0^1 p(0|N, \theta^{\hat{k}+1}) p(\theta|y_N, \hat{k}+1) d\theta && (3.35)\\
=& 1 - \int_0^1 p(0|N, \theta^{\hat{k}+1}) \text{Be(A,B)} d\theta && (3.36)\\
=& 1 - \frac{1}{\beta(A,B)} \int_0^1 (1-\theta^{\hat{k}+1})^N \theta^A (1-\theta)^B d\theta, && (3.37)
\end{aligned}$$

where A and B are calculated from observed distribution as $A = \sum_{i=1}^N y_i + a - 1$ and $B = \sum_{i=1}^N (\hat{k}+1) - y_i + b - 1$. In the absence of a simple form of the integral in equation 3.37, we turn to Monte Carlo integration. Calculation of α entails integrating a function over a probability distribution. In particular, integration is over the conditional posterior of θ, ie.: $\int_0^1 f(\theta) p(\theta|y_N, \hat{k}+1) d\theta$, where $f(\theta)$ is the probability of observing zero events of size $\hat{k}$ +1 under the null hypothesis. If we can draw independent and identically distributed (iid) samples from a probability distribution, then a finite number of such samples can be used to approximate the integration. For example, if we draw S samples $\tilde{\theta}$ from the distribution $p(\theta)$, then$\int f(\theta)p(\theta)d\theta \approx \frac{1}{S}\sum_{i=1}^S f(\tilde{\theta}_i)$. Fortunately, the form of the conditional posterior of θ is simple (equation 3.17), so generating iid samples, $\tilde{\theta}$, can be achieved by drawing beta random variables : $\tilde{\theta} \sim$ Be(A,B). Then α can be estimated as,

$$\alpha \approx 1 - \frac{1}{S}\sum_{i=1}^{S} p(0|N, (\tilde{\theta}_i)^{\hat{k}+1}) \tag{3.38}$$

$$= 1 - \frac{1}{S}\sum_{i=1}^{S}(1 - (\tilde{\theta}_i)^{\hat{k}+1})^N. \tag{3.39}$$

In this way, a proper estimate of confidence can be calculated which takes into account the total uncertainty in all the model parameters. The Monte Carlo integral in equation 3.39 converges quickly as is shown Figure 3.6A which is a plot of the estimate of α (for some simulated dataset) as a function of the numbers of samples, $\tilde{\theta}$. In Figure 3.6B is a comparison of the estimate of confidence as calculated using the two methods developed so far. The curve is a plot of confidence as a function of sample size for a point estimate of θ ($\alpha_{\hat{\theta}}$). The circles the corresponding estimate of confidence using the Bayesian model (α_B) which takes into account the full uncertainty in θ. Generally, the simplified estimation of confidence ($\alpha_{\hat{\theta}}$) overestimates confidence due to the assumption that θ is known with certainty. Indeed, the estimated confidence when considering the full parameter uncertainty is consistently less than with the simplified approach and thus this method affords a more realistic and conservative estimate of parameter confidence. Finally, note that as N increases, so too will A and B (of equation 3.17). The result is that the

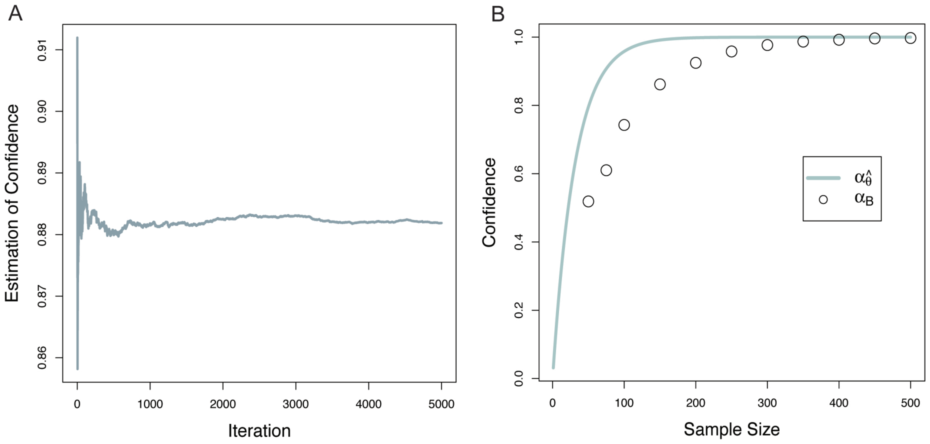

Figure 3.6: Estimation of parameter confidence using Monte Carlo integration. (A) The estimate of confidence from equation 3.39 as a function of the number of posterior samples, $\tilde{\theta}$, used for Monte Carlo integration. This estimate converges quickly. (B) Confidence as a function of sample size estimated using a point estimate of θ ($\alpha_{\hat{\theta}}$) and the Bayesian estimate (α_B). It is clear that $\alpha_{\hat{\theta}}$ overestimates parameter confidence and that the true uncertainty in θ must be taken into account for a accurate estimate of confidence.

conditional posterior of θ becomes narrowed and more probability mass is located closer to the optimal estimate, $\hat{\theta}$ (see Figure 3.2). In the limit of large N, the estimate of θ shrinks to a points estimate and confidence calculated via equation 3.34 converges exactly to equation 3.29 (see also Figure 3.6B). Thus, the Bayesian method presented here is a generalized approach which collapses to the more simplified estimate in the limit of large sample sizes.

I now address a related problem when interpreting such data, which is discussed only briefly as the basic method was proposed previously in (Groulx et al., 2011). Consider that an imperfect data collection algorithm induces artifactual observations into the distribution. In particular, suppose that the largest number of observed bleaching steps, $\hat{k}$, occurs with an anomalously low prevalence and we are tempted to conclude that all $y_i = \hat{k}$ are artifactual and the true n is $\hat{k} - 1$. Given that we have observed K events of size $\hat{k}$, we simply need to consider the sampling distribution of events of size $\hat{k}$ under the hypothesis Bn($\hat{k}$, θ), and calculate the rarity of K in this sampling distribution. As before, this sampling distribution is binomial, Bn(N,$\theta^{\hat{k}}$), and we simply calculate $p(K|N, \theta^{\hat{k}})$. Previous authors estimated this sampling distribution using the Poisson approximation to the binomial distribution and using a fixed estimate of θ (see supplemental of (Groulx et al., 2011)). We generalize this by considering the full uncertainty in the estimate of θ. The probability of observing K or fewer events of size $\hat{k}$ under the null model Bn$(\hat{k}, \theta)$ will be denoted γ. If we integrate over the uncertainty in θ, then γ is calculated as,

$$\gamma = \int_0^1 \sum_{j=1}^{K} p(j|N, \theta^{\hat{k}}) p(\theta|y_N, \hat{k}) d\theta \tag{3.40}$$

$$\approx \frac{1}{S} \sum_{i=1}^{S} \sum_{j=1}^{K} p(j|N, (\tilde{\theta}_i)^{\hat{k}}) \tag{3.41}$$

$$= \frac{1}{S} \sum_{i=1}^{S} \sum_{j=1}^{K} \frac{N!}{(N-j)!j!} ((\tilde{\theta}_i)^{\hat{k}})^j (1 - ((\tilde{\theta}_i)^{\hat{k}}))^{N-j}. \tag{3.42}$$

Here, we have again drawn samples, $\tilde{\theta}$, from the posterior of θ to use for Monte Carlo integration. This integration is approximated by the sum over i in the above equation. The rest of equation 3.42 is the sampling distribution of observations of size $\hat{k}$ and we sum up to K to calculate the probability of seeing K or fewer observations. If γ is very small, it means that our observation of K instances of size $\hat{k}$ is quite rare under the model $\text{Bn}(\hat{k}, \theta)$ and that we might exclude all observations of size $\hat{k}$ as artifacts and accept the hypothesis $n = \hat{k} - 1$.

A potential complication that has been ignored in this work is the possibility of multiple complexes within the same observation volume. This could occur if the density of complexes is sufficiently high, or if complexes have a tendency to cluster together. In this instance, the observed distribution of bleaching events would be drawn from a heterogenous population of species, some of which contain n subunits and others which contain some multiple of n subunits. In fact, this complication seems to be fairly common in the literature and the interpretation of such artifactual data needs to be formally addressed.

In previous work, the strategy of fitting sums of binomial distributions proved successful at overcoming this complication (McGuire et al., 2012). This strategy would be useful only to the extent that the uniqueness of fits could be established. In principle, the methods presented in this paper could be generalized to a model of heterogeneous populations of binomially distributioned observations. Such a model would necessarily have more parameters which would exacerbate the problem of ill-posed inference. However, these methods of confidence estimation should be applicable to a more generalized model. Future work remains to be done in this area.

3.3 Conclusions

Single molecule photobleaching is a pervasive tool for determining protein association which relies on attaching fluorescent probes to molecules of interest and counting distinct photobleaching events. Since there is a non-zero probability of not observing a particular fluorophore, the resulting distribution of photobleaching steps will be binomial. While it seems a straightforward task to interpret such data and deduce stoichiometry, I show that this inference is ill-posed. This means that many possible combinations of n and θ can produce very similar observations. Since there is not generally a unique and optimal estimate of the relevant parameters for a given dataset, extracting the stoichiometry can be error-prone without careful analysis. I develop a general inference model for this type of data which takes into account the full uncertainty in all model parameters. Using this framework, I develop methods for hypothesis testing and calculating parameter confidence which allows for a rigorous interpretation of such data. This work provides a rigorous analytical basis for the interpretation of single molecule photobleaching experiments.

Chapter 4

Modeling Single Molecule Time Series with Nonparametric Bayesian Inference

Abstract The ability to measure the properties of proteins at the single molecule levels offers an unparalleled glimpse into biological systems at the molecular scale. The interpretation of single molecule time series has often been rooted in statistical mechanics and the theory of Markov processes. While these methods have been helpful, they are not without significant limitations including problems of model selection and parameter non-identifiability. To overcome these challenges, I introduce the use of nonparametric Bayesian inference for the analysis of single molecule time series. These methods provide a flexible way to extract structure from data instead of assuming models beforehand. I demonstrate these methods with applications to several diverse settings in single molecule biophysics. These methods bring a more powerful approach to the study of molecular biophysics.

4.1 Introduction

Proteins are the fundamental unit of computation and signal processing in biological systems. Understanding the biophysical mechanisms that underlie protein conformational change remains an important challenge in the study of biological systems.The ability to measure the properties of proteins at the single molecule levels offers an unparalleled glimpse into biological systems at the molecular scale. The was first achieved with ion channel proteins using the patch clamp technique (Hamill et al., 1981) and has been extended to soluble proteins using optical methods such as single molecule FRET (Weiss, 2000) and optical tweezers (Svaboda et al., 1993). Such single molecule time series reveal stochastic dynamics indicative of rapid transitions between semi-stable conformational states separated by free-energy barriers. This leads to a natural interpretation of these time series within the context of equilibrium statistical physics and the theory of Markov processes. Markov models fit well within the conceptual framework of protein conformational change, yielding mechanistic models with a finite number of discrete energetic states. Even early investigators imagined that proteins achieve their functions by accessing a small number of physical states (Hodgkin and Huxley, 1952), a framework that persists today. In practice, single molecule time series, inevitably obscured by experimental noise and other obfuscations, are often analyzed using hidden Markov models (Rabiner, 1989). While this approach has been widely successful, it is not without important limitations.

Current methods for the analysis of single molecule time series suffer from problems of model selection and parameter identifiability. Analysis of single molecule time series often begins with the investigator positing some mechanistic model: the lens through which the data are to be interpreted. Generally, we must postulate the existence of some number of biophysically relevant states and perhaps even their interrelationships. For example, in the study of ion channel gating, a typical analysis requires postulating a particular mechanistic scheme consisting of a specified state space and connectivity, and using a maximum likelihood approach to estimate the relevant parameters of that scheme, given the data (Colquhoun and Hawkes, 1981; Horn and Lange, 1983; Qin et al., 1997). However, in most applications, the number of hidden states is not obvious from the data and is not known beforehand. In fact, it is likely the case that an experiment was performed with the purpose of uncovering the existence and details of hidden molecular states. Therefore, the choice of a particular model has a very strong effect on the analysis and interpretation of the data. Methods for aiding in this problem of model selection have been proposed for a variety of experimental settings and generally have relied on model comparison via maximizing likelihood or penalized maximum likelihood such as Akaike Information Criterion (Horn, 1987; Ball and Sansom, 1989; Liebovitch and Toth, 1990; Wagner and Timmer, 2001; Csanady, 2006). The strategy with such methods is motivated by parsimony: the goal is to find the model which provides the best explanation of the data, yet remains the least complex. While parsimony is likely a useful guiding principal, these methods leave us with no rigorous way of quantifying our confidence in models relative to each other; we must rely on ad hoc comparison of models based on AIC score. Additionally, maximum likelihood methods are generally unable to detect parameter non-identifiability, a pitfall that is increasingly common as researchers pursue models of higher complexity (Siekmann, et al., 2012; Calderhead et al., 2013; Hines et al., 2014). Though many attempts have been made, likelihood-based approaches for modeling single molecule time series have proven inadequate.

Here I introduce a novel approach for the analysis of single molecule time series which circumvents the problem of model selection by using non-parametric Bayesian inference. The goal of these methods is to use a class of probability models which are so flexible that we are able to *extract structure* from data instead of assuming models beforehand. These methods have become widely used in the machine learning community to handle challenging problems such as document modeling (Blei at al, 2004), speaker diarization (Fox et al., 2011), and image processing (Kivinen et al., 2007), among many others. The approach relies on the theory of Random Probability Measures and in particular, I use the Dirichlet process to provide an infinite dimensional probability model with well-defined properties for modeling finite data. This infinite model subsumes the set of all possible models, but in fitting finite data, we learn which of the infinite model components are actually necessary to provide a good explanation of the data. The properties of Dirichlet process models yield parsimony while preventing over fitting, allowing us to *discover* what process generated the data, instead of assuming it. Importantly, this Bayesian approach provides estimates of all parameters, their uncertainty, and their identifiability. Finally, because our infinite dimensional model is a well-defined probability distribution with well-known properties, we gain a quantification of parameter confidence for different models which can be used for model comparison.

I demonstrate the use of nonparametric Bayesian inference with three use cases from single molecule biophysics. Using a Dirichlet process mixture model, I show that dwell-times from single ion channel recordings can be modeled nonparametrically in order to discover the number of biophysical states hidden in the data. I then describe the hierarchical Dirichlet process hidden Markov model and apply this model to time series from electrophysiology, single molecule photobleaching and single molecule FRET. Finally, I introduce the hierarchical Dirichlet process aggregated Markov model which allows us to nonparametrically analyze single ion channel recordings and extract open and closed states without specifying a model. These methods provide a flexible and powerful framework for the analysis of diverse types of single molecule data.

4.2 Methods

Electrophysiology

HEK293 cells were cultured following standard protocols. Wild-type BK channel cDNA was transiently transfected into HEK cells with Lipofectamine 2000. As an optical marker, Enhanced green fluorescent protein (EGFP) was cotransfected. Recordings of single BK channels were performed 2-4 days after transient transfection. Voltage-clamp was performed on inside-out patches pulled from HEK cells at room temperature. Patch electrodes were 1-2 MΩ and electrode solution contained (in mM): 6 KCl, 136 KOH, 20 Hepes, 2 $MgCl_2$, and pH adjusted to 7.2 using $MeSO_3H$. Bath solution contained (in mM): 6 KCl, 136 KOH, 20 HEPES, .01 Crown Ether, and pH was adjusted to 7.2 using $MeSO_3H$. Additionally, EGTA was added to buffer calcium and varying amounts of $CaCl_2$ were added. Free calcium concentrations were measured using a calcium-sensitive electrode. Recordings were performed with an Axopatch 200A amplifier and digitized using an ITC-16 A/D converter. Single channel traces were sampled at 100 kHz and analog filtered at 10 kHz, and collected using PatchMaster software.

FRET

The single molecule FRET data from NMDA receptor proteins were kindly contributed by David Cooper and Christy Landes at Rice University and was collected using procedures similar to those reported in (Ramaswamy et al. 2012). The agonist-binding domain of the NMDA receptor was expressed and purified using standard procedures. Streptavidin acted as a linker between a biotin-PEG slide and the biotin-conjugated anti-histidine antibody bound to the NMDA subunit. A PBS solution containing 250 nM protein tagged with biotin-conjugated anti-histidine monoclonal antibody was then added. To obtain the smFRET trajectories for the individual protein molecules, a 10 x 10-μm area of the sample was scanned to spatially locate 20-25 molecules. The fluorescence signals of the donor and the acceptor were collected until the fluorophores were photobleached. Photon counts were collected from two APDs tuned to the wavelengths for acceptor and donor light which were then processed to remove background signal and crosstalk from the signals and FRET efficiency was calculated using standard methods. The emission intensity trajectories were collected at 1-ms resolution and later binned to 10-ms time steps.

Photobleaching

The single molecule photobleaching data was kindly contributed by John Bankston at the University of Washington. The data were collected as described in (Bankston et al., 2012), which is briefly replicated here: Oocytes were injected with varying ratios of TRIP8b and HCN2 mRNA. TIRF movies were acquired using a Nikon TE2000-E microscope with a high numerical aperture objective (100x, 1.49 N.A.; Nikon) and the Evolve 512 EMCCD camera (Photometrics). Oocytes were illuminated with a 488-nm argon laser from Spectra Physics. An image stack of 800 - 1200 frames was acquired at 30 - 50 Hz. The first five frames after opening of the laser shutter were averaged, and the background was subtracted using the rolling-ball method in Image J (National Institutes of Health). The image was then lowpass-filtered with a 2-pixel cutoff, and thresholding was applied to find connected regions of pixels that were above threshold. A region of interest (ROI) of 6 x 6 pixels was placed around the center of the spot. Spots smaller than 3 pixels and larger than 15 pixels were discarded manually. Finally, the summed fluorescence intensity inside the 6 x 6 ROI was measured and plotted the data versus time.

Data Analysis

The models and algorithms used for data analysis are described in detail in the next section. Analysis for the Dirichlet process mixture of exponential was performed using scripts written in R. For the hierarchical Dirichlet process hidden Markov model, the beam sampling implementation of (Van Gael at al., 2008) was used (code available at mloss.org/software/view/205/). For the hierarchical Dirichlet process aggregated Markov model, scripts were written in Matlab.

4.3 Theory

4.3.1 Nonparametric Bayes

Methods of nonparametric Bayesian inference rely on a class of flexible probability distributions known as random probability measures (RPM). While there is an extensive literature on RPMs of all flavors, I focus on the Dirichlet process (Hjort et al., 2010; Mueller and Rodriguez, 2012). The Dirichlet process, DP(α, H), is a distribution on distributions (Ferguson, 1973). It has two parameters: a scalar α which is referred to as the concentration parameter, and a base distribution H. Draws from DP(α, H) are random probability measures which are centered on H and whose variance about H is controlled by α. A useful representation of a draw from a Dirichlet Process is the *stick-breaking* process of (Sethuraman, 1994). A random probability measure, G, is drawn from a Dirichlet Process as follows,

$$G \sim DP(\alpha, H) \tag{4.1}$$

$$G = \sum_{i=1}^{\infty} w_i \delta_{\theta_i}, \tag{4.2}$$

where all θ_i are independent and identically distributed (iid) samples from the base distribution H, and the weights satisfy the following stick-breaking construction: $w_i = v_i \prod_{k<i}(1 - v_k)$ for $v_k \sim \text{Beta}(1, \alpha)$. Imagine breaking a stick of unit length into an infinite number of segments in the following way. Break the stick at a random location $w_1 \sim \text{Beta}(1, \alpha)$ and associate with this weight a random draw from H, $\theta_1 \sim H$. The remaining

length of the stick is now $1 - w_1$. Again draw $v_2 \sim \mathrm{Beta}(1, \alpha)$ and break the remaining length of the stick at this location, such that $w_2 = (1 - w_1)v_2$ and associated with this weight is another iid draw from the base measure, $\theta_2 \sim H$. This process is repeated infinitely with the result that the probability mass is distributed across a countably infinite number of segments. For convenience, I denote the sequence $w_1, w_2, w_3, ...$ which satisfies the stick-breaking construction as $w \sim \mathrm{GEM}(\alpha)$ (Pitman, 2002). The expectation of the size of each weight, $E[w_i]$, decreases geometrically with i, such that only finitely many w_i occupy nearly all the probability mass while the infinitely many others occupy negligible probability. From equation (4.2), we see that G is an infinite mixture of components each with probability mass w_i located at θ_i (see Figure 4.1). Note that G is a discrete probability distribution, even though H might be continuous.

4.3.2 Dirichlet process mixture models

Since a draw from the Dirichlet Process is discrete, it can be awkward when used with data known to be drawn from a continuous distribution. A common variation is a Dirichlet Process mixture model (DPMM), where G is convolved with some parametric continuous distribution (Lo, 1984). Since G is a discrete distribution, this convolution results in a mixture model with an infinite number of components. The data y_i are drawn from a DPMM as,

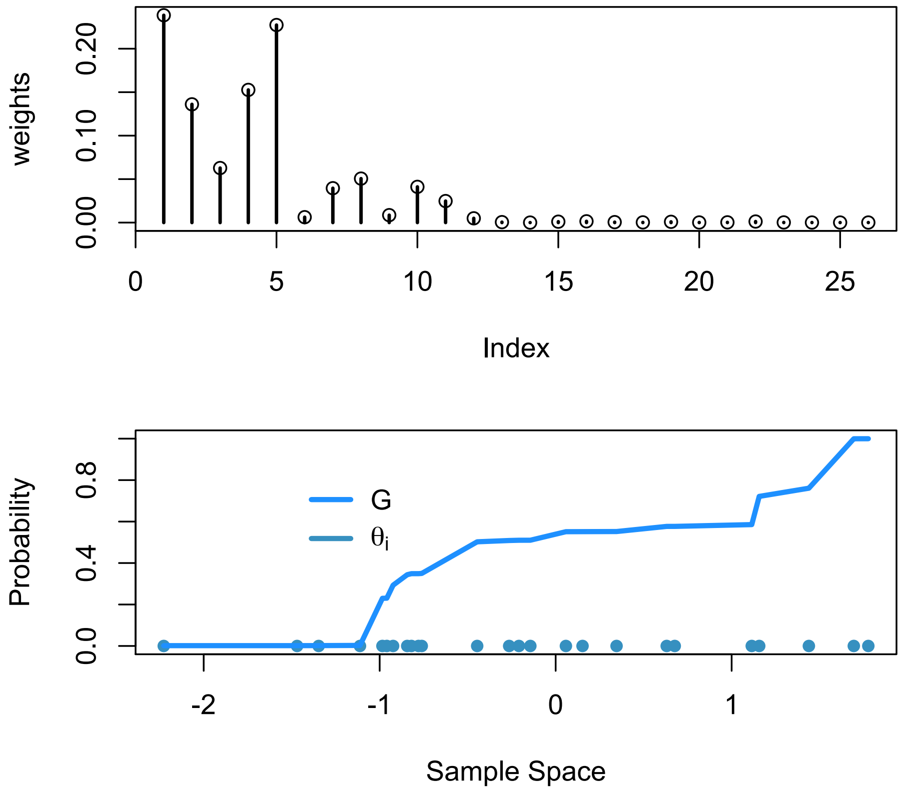

Figure 4.1: An example draw from a Dirichlet process, $G \sim DP(\alpha, H)$, with only finitely many (25) components of G visualized. (Top) Stick-breaking weights, w_i: this infinite sequence of weights sums to 1, yet most of the probability mass is occupied by finitely many of them. (Bottom) All θ_i are drawn iid from the base measure, which is not visualized here but was a Normal distribution centered at 0 with unit variance. The distribution G is discrete, with point masses w_i located at θ_i. For visualization, the cumulative sum of G is shown.

$$G \sim DP(\alpha, H) \tag{4.3}$$

$$y_i \sim \int p(y_i|\theta)G(d\theta) \tag{4.4}$$

$$= \sum_{j=1}^{\infty} w_j p(y|\theta_j) \tag{4.5}$$

We now imagine that each data point is drawn from one of an infinite number of clusters, each one parameterized by θ_j. Due to the properties of the stick-breaking process, only finitely many w_j occupy nearly all the probability mass, while infinitely many others occupy negligible probability mass. Since the data y_i are sampled from the probabilities w_j, a natural clustering is induced in the data. In principle, the number of inferred clusters could range between two extremes: there could be one cluster from which all the data are drawn, or there could be N clusters, each data point being drawn from its own component. Obviously, neither of these of these extremes is particularly useful. Most commonly, we infer with high posterior probability the presence of some small number of clusters k^*, where $k^* << N$.

As an example, later I will model dwell-times from single ion channel recordings using a mixture of exponential distributions. In this case, $p(y|\theta)$ is an exponential distribution with unknown scale parameter θ. If we do not know how many clusters (mixtures) are in the data, we can use DPMM to model an infinite mixture

$$y_i \sim \sum_{j=1}^{\infty} w_j e^{-(\theta_j/y)}. \tag{4.6}$$

Model Inference

Inference with this infinite mixture model is achieved with the following Gibbs Sampling scheme. I first describe the relevant conditional posterior distributions for sampling a finite mixture of exponential distributions, and then how sampling is performed with the infinite mixture model. For a given finite mixture model with K components, we are interested in computing the marginal posterior distributions of $\theta_1, \theta_2, ..., \theta_K$. The likelihood is an exponential distribution,

$$\begin{aligned} p(y_i|...) &\propto w_1 e^{-(\theta_1/y)} + w_2 e^{-(\theta_2/y)} + ... + w_K e^{-(\theta_K/y)} && (4.7) \\ &= \sum_{j=1}^{K} w_j e^{-(\theta_j/y)} && (4.8) \end{aligned}$$

For the prior on θ I use a conjugate gamma distribution, $\text{Ga}(A, B)$. For a single-component exponential distribution with a gamma prior, the posterior distribution of scale parameter θ is,

$$p(\theta|y_N) \propto \prod_{i=1}^{N} we^{-(\theta/y_i)}\text{Ga}(A,B) \tag{4.9}$$

$$= \prod_{i=1}^{N} we^{-(\theta/y_i)}\frac{B^A}{\Gamma(A)}\theta^{A-1}e^{(-B\theta)} \tag{4.10}$$

$$\propto \theta^{(A+N)}e^{\sum y_i+B} \tag{4.11}$$

$$= \text{Ga}(A+N, B+\sum y_i) \tag{4.12}$$

For the mixture model, I introduce a latent indicator variable, s_i, which serves to label each data point according to which component it was likely drawn from. Using these indicator variables, the posterior of θ is extended to multiple components. Let A_j be set of all i such that $s_i = j$. Then the posterior over θ_j goes as,

$$p(\theta_j|y_N, s_1, ..., s_N) \propto \prod_{i\in A_j} we^{-(\theta/y_i)}\text{Ga}(A,B) \tag{4.13}$$

$$= \text{Ga}(A+|A_j|, B+\sum_{i\in A_j} y_i) \tag{4.14}$$

For each s_i, we sample the conditional posterior of datapoint y_i belonging to each of the K components from a Multinomial distribution,

$$p(s_i = j|...) \propto w_j p(y_i|\theta_j) \tag{4.15}$$

$$p(s_i|...) \propto \text{Mult}(p(s_i = 1|...), p(s_i = 2|...), ..., p(s_i = K|...)). \tag{4.16}$$

The cluster weights, w_j, are drawn from the standard Dirichlet distribution,

$$p(w_1, w_2, ..., w_K|...) \propto \text{Dir}(|A_1|, |A_2|, ..., |A_K|). \tag{4.17}$$

For any mixture model with K components, the conditional posterior distributions described above (equations 4.14, 4.16, 4.17) specify an efficient Gibbs sampler for calculating the posterior distributions of all model parameters. However, we aim to rely on the properties of the Dirichlet process in order to model an infinite number of clusters. The only change to the previous Gibbs sampler is how to deal with an infinite number of clusters. Sampling s_i from a Multinomial distribution with K components is likely to be impossible as $K \to \infty$. Recall, however, that the Dirichlet process has the useful property that finitely many components occupy most of the probability mass while the infinitely many other one occupy a negligible amount. Thus, even though $\theta_1, \theta_2, ...$ is infinitely long, the DP induces a natural clustering such that the data are drawn from a finite set, θ^*. During any particular iteration of Gibbs sampling, let k^- denote the number of components currently represented in θ^*. Then data point y_i might be sampled from one of the k^- clusters which are already represented, or from one of the infinitely many other clusters which are not yet represented, but all of whom together occupy finite probability mass. Sampling the indicator variables s_i goes as,

$$p(s_i = j|\mathbf{s}^-, \mathbf{y_N}) \propto \left\{ \begin{array}{rl} n_j \int p(y_i|\theta_j^{*-}) dp(\theta_j^{*-}|y_j^{*-}) & j \leq k^- \\ \alpha \int p(y_i|\theta_j) dG(\theta_j) & j > k^- \end{array} \right. \tag{4.18}$$

Thus, the indicator variables sample from each existing component with probability proportional to the current size of the component, and generate a new component with probability proportional to α. If we use a conjugate model, then computing the integrals in equation 4.18 is simple and this scheme can be used for sampling from an infinite number of clusters. In this case of DP mixture of exponentials, we indeed can utilize the conjugacy between exponential and gamma distributions and the previous method can be used for inference. Alternatively, I prefer to use the slice sampling method of (Walker, 2007) because it's clever as hell.

Recall that, generally, our mixture model posits that the data are drawn from an infinite mixture of parametric distributions,

$$p(y_i|...) = \sum_{j=1}^{\infty} w_j p(y|\theta_j). \tag{4.19}$$

We augment this model by adding a latent variable u, drawn from a uniform distribution, such that the joint model is,

$$p(y_i, u|...) = \sum_{j=1}^{\infty} \mathrm{I}(u < w_j) p(y|\theta_j). \tag{4.20}$$

Note that marginalization of equation 4.20 with respect to u results in the original model (equation 4.19), ie. $\int p(y_i, u|...)du = p(y_i|...)$. Since all w_j are less than one, any particular draw of u partitions the infinite set of w_j into two sets: a finite set for which $w_j > u$ and an infinite set for which $w_j < u$. By incorporating this augmented model into the Gibbs sampler, we can sample u in order to only represent finitely many clusters at each iteration, yet the aggregate sampling marginalizes the model back to that of equation 4.19. For each iteration, we draw $u_1, ..., u_N$ uniformly on the intervals $(0, w_{s_i})$ and represent k^* clusters where

$$\sum_{j=1}^{k^*} w_j > 1 - \min(u_1, ..., u_N). \tag{4.21}$$

Each iteration is simply a finite mixture model and the number of mixture components fluctuates over the course of MCMC to sample the infinite number of clusters.

Demonstration

Figure 4.2 shows an example of using this infinite mixture model. At top left, a simulated dataset was drawn from a mixture of four exponential distributions. The four components had scale parameters, θ_j, equal to {.001,.01,.1,10}. Data points are plotted logarithmically to aid in visualization and a kernel density estimate is shown. When shown in this way, we might guess by eye that there are distinct clusters in the data, but we would be unsure of how many. Using a DP mixture of exponentials allows us to posit

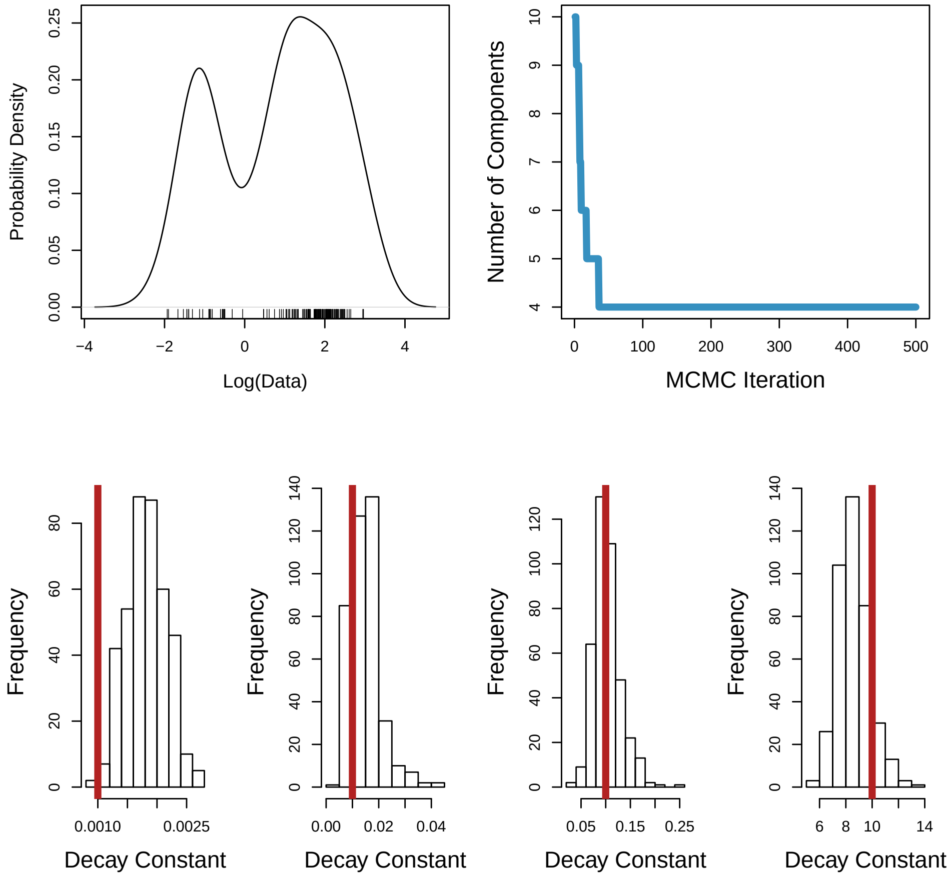

Figure 4.2: Demonstration of Dirichlet process mixture of exponentials. (Top left) Simulated data drawn from a mixture of 4 exponential distributions with scale parameters, θ_j, equal to $\{.001,.01,.1,10\}$, plotted logarithmically. (Top right) Result of modeling this dataset with DP mixture of exponentials: the infinite model converges to 4 components. (Bottom) Marginal posterior distributions of all θ_j that remain in the model. True values shown as red vertical lines. Algorithm parameters: $\alpha = 1$.

an infinite model and then learn how many clusters are actually in the data. Shown at top right is the number of components in the infinite model that are represented throughout the course of MCMC simulation. The model initializes with an arbitrary, large number of clusters, but quickly converges to the correct number. The bottom row of Figure 4.2 shows the marginal posterior distributions for each of the θ_j that remain in the model. The true values of each θ_j are shown as red vertical lines. Using the DP mixture of exponentials, we were able to correctly learn the number of clusters in the data, and also get an accurate quantification of the relevant model parameters and their uncertainty.

4.3.3 Infinite Hidden Markov Model

Hidden Markov models (HMMs) have enjoyed vast application in many areas of science and engineering due to their flexibility and predictive ability. In this model, it is assumed that observable data, y_t, is an obfuscation of a hidden dynamical process that we cannot directly access. In particular, it is assumed that the system of interest has access to K different hidden states $(1, 2, ..., K)$ and transitions stochastically between states at every time step. The dynamics of the system are fully captured by the transition probability matrix π, where each element $\pi_{i,j}$ is equal to $p(s_t = j|s_{t-1} = i)$, the probability of a transitions to state j from state i at each time step. Atop these dynamics, it is assumed that each hidden state, s, has a distinct emission distribution, $p(y_t|s_t)$. Therefore, the system transitions stochastically according to π, and each observation is a random draw from $p(y_t|s_t)$. Because of the Markov property, the joint probability of all hidden states and observations can be written as

$$p(s_N, y_N|...) = \prod_{t=1}^{N} p(s_t|s_{t-1})p(y_t|s_t). \tag{4.22}$$

While HMMs have been widely useful, a major limitation is that we must specify how many hidden states, K, are in the model. To overcome this barrier, (Beal et al., 2002) introduced the infinite hidden Markov model (iHMM) which was later generalized and termed the hierarchical Dirichlet process hidden Markov model (Teh et al., 2006; Fox et al., 2008). In this model, the number of hidden states is left unknown, and the transition matrix, π, is modeled nonparametrically using the hierarchical Dirichlet process. Each row of π is a draw from a Dirichlet process and thus specifies the probability of transitioning to each of an infinite number of other hidden states. In order to ensure the all the rows of π are coupled, each row π_i is drawn from a DP with base distribution β, which itself is a draw from a Dirichlet process,

$$\beta \sim \text{GEM}(\gamma) \tag{4.23}$$

$$\pi_i \sim \text{DP}(\alpha, \beta) \tag{4.24}$$

$$\theta_i \sim H \tag{4.25}$$

$$y_t \sim p(y_t|s_{\theta_i}). \tag{4.26}$$

The goal is then to learn the number of hidden states from a particular time series.

Model Inference

I first describe a Gibbs sampling scheme for parameter inference with finite HMMs and then describe the implementation I use for the iHMM. For these examples, we imagine our observations are normally distributed random variables and that each hidden state corresponds to a distinct mean θ_i and precision τ_i, such that $y_t \sim N(\theta_i, \frac{1}{\tau_i})$. Again, let A_i denote the set of all t for which $s_t = i$. For the means, θ_i, I use a conjugate prior normal distribution $N(a, b)$. For each θ_i,

$$\begin{aligned}
p(\theta_i|...) &\propto N(M, V) && (4.27)\\
\text{where} \quad M &= \frac{ab + \tau \sum_{t\in A_i} y_t}{|A_i|\tau + b} && (4.28)\\
V &= \frac{1}{|A_i|\tau + b} && (4.29)
\end{aligned}$$

With a conjugate gamma prior,$p(\tau_i) = \text{Ga}(c, d)$, on the precisions, τ_i,

$$\begin{aligned}
p(\tau_i|...) &\propto \text{Ga}(A, B) && (4.30)\\
\text{where} \quad A &= \frac{d + |A_i|}{2} && (4.31)\\
B &= \frac{1}{bc + \frac{1}{2}\sum(y_t - \theta_i)^2}. && (4.32)
\end{aligned}$$

Sampling the transition matrix, π, is simple conditioned on the previous samples of hidden states $s_1, ..., s_N$. First, we use the standard Dirichlet distribution prior for rows of the transition matrix, ie. $p(\pi_i) = \text{Dir}(m, ..., m)$. Let matrix N track the number of transitions between hidden states i and j such that $N_{i,j} = \sum_t I(s_t = j | s_{t-1} = i)$. Then each row of the transition matrix is sampled as,

$$p(\pi_i|...) \propto \text{Dir}(N_{i,1} + m, ..., N_{i,K} + m). \tag{4.33}$$

Finally, the hidden states, s_t, are sampled using the forward-filter-backward-sampler method (Scott, 2002). First we construct the $K \times N$ forward matrix F in the following way. For each datapoint, y_t, first compute vector O which quantifies the conditional probability of observing y_t given the emission distributions of each hidden state,

$$O = \begin{bmatrix} p(y_t|\theta_1, \tau_1) \\ p(y_t|\theta_2, \tau_2) \\ . \\ . \\ . \\ p(y_t|\theta_K, \tau_K) \end{bmatrix}. \tag{4.34}$$

We then combine the observation probabilities, the transition probabilities, and the occupancy probabilities from the previous time step,

$$L = (O \times \pi) \bullet F_{,t-1} \tag{4.35}$$

$$F_{,t} = \frac{L}{\sum L}. \tag{4.36}$$

Having computed F deterministically, we use Gibbs sampling on the backwards pass. Starting at time step N, we move backwards through each time step t, and combine F with the transition probability

$$L = F_{,t} \bullet \pi_{,s_{t+1}} \tag{4.37}$$

$$\vec{p} = \frac{L}{\sum L}. \tag{4.38}$$

We sample s_t from the resulting multinomial distribution,

$$p(s_t|...) \propto \text{Mult}(\vec{p}). \tag{4.39}$$

The result of this forward-backward sampler is a new sample of $s_1, s_2, ..., s_N$. For any hidden Markov model of fixed size, K, this Gibbs sampler allows us to calculate posterior distributions of all relevant parameters.

Generalizing this model to the infinite case will proceed similarly as with the mixture model. Again, the problem is that we now wish to consider the probability of transitions to each of an infinite number of hidden states, a computation that we cannot perform in our existing Gibbs sampler. However, using the hierarchical Dirichlet process hidden Markov model, we can sample from both the currently instantiated hidden states as well as the infinitely many other hidden states which have yet to be sampled (Teh et al., 2006),

$$p(s_t = j|\mathbf{s}^{-}, \beta, \alpha, \mathbf{y_N}) \propto \begin{cases} (N_{s_{t-1},j} + \alpha\beta_j)\frac{N_{s_{t+1}}+\alpha\beta_{s_{t+1}}}{N_{k,}+\alpha} & j \leq k^{-}, k^{-} \neq s_{t-1} \\ (N_{s_{t-1},j} + \alpha\beta_j)\frac{N_{s_{t+1}}+1+\alpha\beta_{s_{t+1}}}{N_{k,}+1+\alpha} & j = s_{t-1} = s_{t+1} \\ (N_{s_{t-1},j} + \alpha\beta_j)\frac{N_{s_{t+1}}+\alpha\beta_{s_{t+1}}}{N_{k,}+1+\alpha} & j = s_{t-1} \neq s_{t+1} \\ \alpha\beta_j\beta_{s_{t+1}} & j = k^{-} + 1 \end{cases} \tag{4.40}$$

The sampling scheme works well, but it was noted that since Markov-type models will inherently have very high correlation between the latent variables, this form of Gibbs sampling could mix very slowly. To remedy this, (Van Gael et al., 2008) proposed the *beam sampler* for iHMMs. This implementation combines the dynamic programming approach described previously (forward-filter backward-sampler) with the slice sampling approach of (Walker, 2008). As described previously, the model is augmented to include latent variables $u_1, ..., u_N$ in order to limit the computation to a finite number of hidden states (at each iteration of MCMC). Once the appropriate number of states, k^*, is computed from $\vec{u}$, then we proceed with the Gibbs sampler just described for finite HMMs. Again, throughout the course of MCMC, resampling $\vec{u}$ results in fluctuations in the number of hidden states represented such that the aggregate of all MCMC samples results in integration over the infinite number of states. Sampling for β is performed using standard sampling methods for hierarchical Dirichlet process models (Teh et al., 2006).

4.3.4 Infinite Aggregated Markov Model

In the iHMM, it was assumed that each hidden state corresponds to a distinct emission distribution, $p(y_t|\theta_i)$. In some cases, we might want to model a degeneracy such that multiple hidden states share the same emission distribution. In this aggregated Markov model (Kienker, 1989), we imagine that the hidden states appear as aggregated into one of A distinct emission distributions such that $A < K$. We augment the iHMM with an indicator variable, $a_t \in \{1, 2, ..., A\}$, that specifies which aggregate each data point is drawn from such that $y_t \sim p(y_t|\theta_{a_t})$. In this case, we cannot identify different states by their emission distributions, but aim to infer the hidden states based on differences in their dynamics. In the next section, this model is applied to data from single ion channel recordings and A is fixed to be two. It is my intention with the iAMM that the number of aggregates, A, is known beforehand and we mean to infer the number of hidden states within each aggregate. I suppose it would be possible to treat the number of aggregates as unknown and model both A and π nonparametrically, but I do not know of any interesting use for such a thing, so I do not explore this possibility.

The use case for the iAMM will be the analysis of single ion channel recordings, for which I add one additional feature to the model. Previous authors extended the infinite hidden Markov model framework by allowing for a strong preference for models with *state-persistence* (Fox et al., 2011). That is, we assume the time-scale of system dynamics is significantly slower than the data sampling rate. In this way, we are interested in solutions to the data where the system stays in each state for many time samples and we are intentionally not interested in models where states have zero dwell-time before transitioning. This certainly seems to be the case with ion channels, where from dwell-time distributions, we imagine that the channel tends to stay in each state for multiple time samples (at least). Following (Fox et al., 2011), I employ a *sticky*-iAMM by biasing probability mass onto the diagonal elements of the transition matrix π. By ensuring non-zero probability mass on the diagonal of π, we exclude models where states transition arbitrarily quickly to other states. To achieve this, I make a slight alteration to the algorithm described in the previous section. We add a hyper-parameter κ, the magnitude of which tunes the *stickiness* of the resulting Markov model. Each row of π is drawn from a Dirichlet process, with the diagonal elements biased by κ,

$$\pi_j \sim \mathrm{DP}(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa}), \tag{4.41}$$

and the rest of the algorithm remains the same. Incorporating uncertainty in κ into the sampling model should be possible in principle (Escobar and West, 1995), but I prefer to use a fixed value. In experiments with simulated data, $\kappa = 100$ works well, and I use this same value for all ion channel data analyzed in the Results section.

An example of the sticky-iAMM, meant to mimic single ion channel recordings, is shown in Figure 4.3. For simulating data, I use $A = 2$ and $K = 4$, and use transition dynamics π such that the two states within each aggregate have very different transition probabilities. A sample of such data is shown at the top of Figure 4.3 and we can even see by eye that within each emission distribution are events that have very long durations and other events with have very brief durations. By using the sticky-iAMM to analyze this time series, we can infer how many states are hidden within the two aggregated states. The result of this model is shown as the colors in the top of Figure 4.3: each datapoint is colored according to which hidden state it likely was drawn from. With the infinite model, we are able to correctly identify that there are four states with distinct dynamics and are able to label all the data points: open states as red and blue and closed states as green and gold. The middle of Figure 4.3 is a plot of the number of hidden states represented throughout the course of MCMC. Figure 4.3 (bottom) shows the posterior distribution over number of hidden states and we see that high posterior probability is placed on there being four hidden states within this time series. Therefore, we are able to accurately infer the number of hidden states within this aggregated Markov process time series.

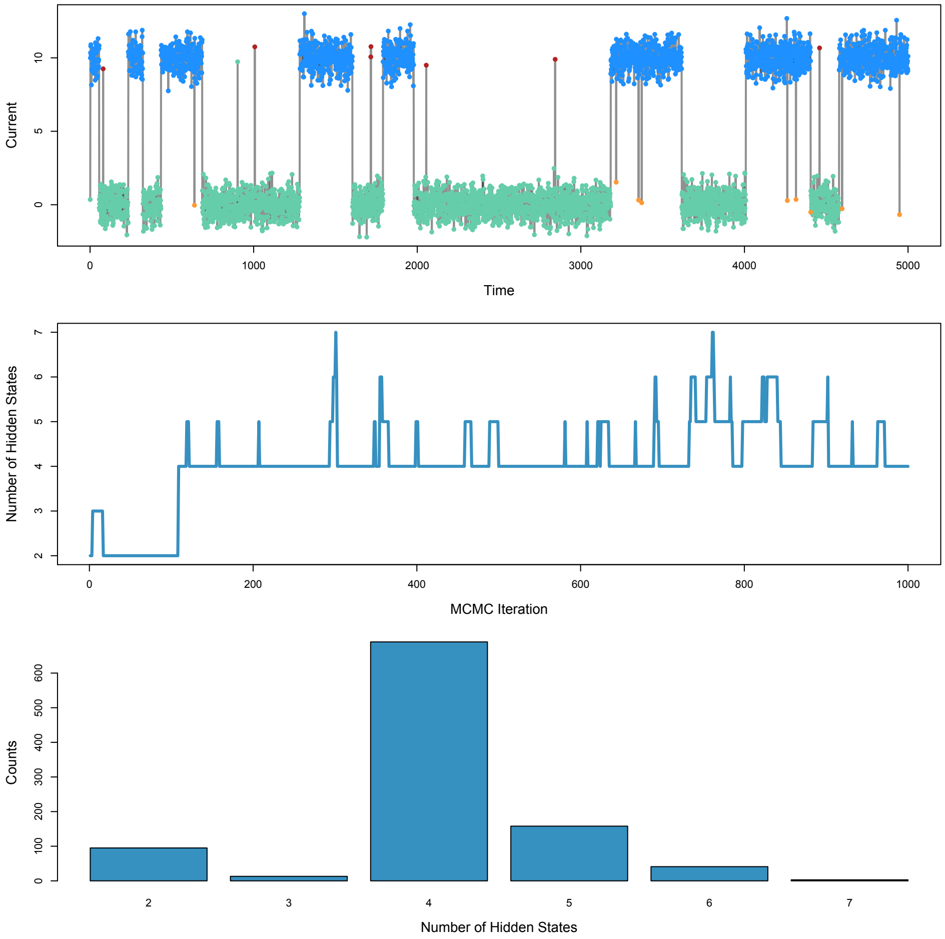

Figure 4.3: Demonstration of the infinite Aggregated Markov Model. (Top) Simulated data from a 4-state process with two closed and two open states with different dynamics. Each of the states differ in their exit rate - we can even tell by eye that there is a short-lived state and a long-lived state, for both open and closed. Colors correspond to the inferred state-assignments when this time series is modeled with the iAMM; we find the number of hidden states correctly and correctly label each data point. (Middle) The number of hidden states over the course of MCMC simulation. (Bottom) The posterior distributions over the number of hidden states. There is high probability that this time series was generated from a 4-state process. Algorithm parameters: $\alpha = 1, \gamma = 1, \kappa = 100$.

Finally, I discuss the effects of Dirichlet process parameters on model inference. Recall that random probability measure, G, is a draw from a Dirichlet process as, $G \sim \mathrm{DP}(\alpha, H)$. The Dirichlet process has two parameters, scalar α and probability measure H. Base measure H serves as the expectation of $G(A)$ (on any interval A) such that $\mathrm{E}[G(A)] = H(A)$. Parameter α alters the variability of G around the expectation H, $\mathrm{Var}[G(A)] = \frac{H(A)(1-H(A))}{\alpha+1}$, such that when α is large, G settles near H with low variance. With respect to the stick-breaking representation of the Dirichlet process, α tunes the expected size of the weights. Since the weights are related to iid draws from a Beta(1,α) distribution, large α results in many weights which are relatively small and a small value of α results in fewer weights which each occupy larger probability mass. Therefore, when using a Dirichlet process prior for model inference, the value of α will have an effect on the number of inferred model components. One approach to handling this complication is to incorporate uncertainty in α into the model by putting a parametric prior on α and marginalizing this uncertainty through the course of MCMC sampling (Escobar and West, 1995). In the applications explored in the next section, I am primarily interested in applying these methods to distinct subsets of data, each of which represents an independent measurement or a measurement in a different experimental condition. In this way, I am most interested in comparing the inference results across different data subsets, where the inference algorithm is fixed in each case. Then, differences between the models inferred from each subset can be meaningfully compared, regardless of the uncertainty in α. Therefore, my strategy for choosing DP parameter values is to choose values which have accurate and reliable performance with simulated data and then fix these parameters for analysis of an entire dataset.

It is important to conduct sensitivity analysis to determine how changes in α affect model inference. As an example, a Dirichlet process mixture of exponentials was used to model data simulated from a mixture of two exponentials where the components differed in time-scale by ten-fold ($N = 200$ data points). Figure 4.4 shows the result of this model inference for several fixed values of α. It is clear that over this range of α, the effect on the inferred models is negligible as the two component mixture is correctly inferred in each case. For the biophysical applications in the Results section, I fix $\alpha = 1$, which, when compared across distinct data subsets, is able to distinguish when a small number of components are in the data. For the Hierarchical Dirichlet process models (iHMM and iAMM), we incur an additional parameter γ, which also tunes the variability of a Dirichlet process around its base measure. Again, I choose to fix $\gamma = 1$, since this low value leads to good performance with simulated data. With the sticky-iAMM, we have an additional parameter κ which biases probability mass onto the diagonal elements of a transition matrix π. I fix $\kappa = 100$, which places a very weak prior on elements of π, since the traces used for analysis have 10^5 data points. Nonetheless, this weak prior is able to deter states which have zero dwell time and effectively accomplishes the goal of the sticky-iAMM. Despite uncertainty in these algorithm parameters, my strategy is to fix them to be small values which perform well with simulated data, because my primary goal is to compare between data sets given fixed values of these parameters.

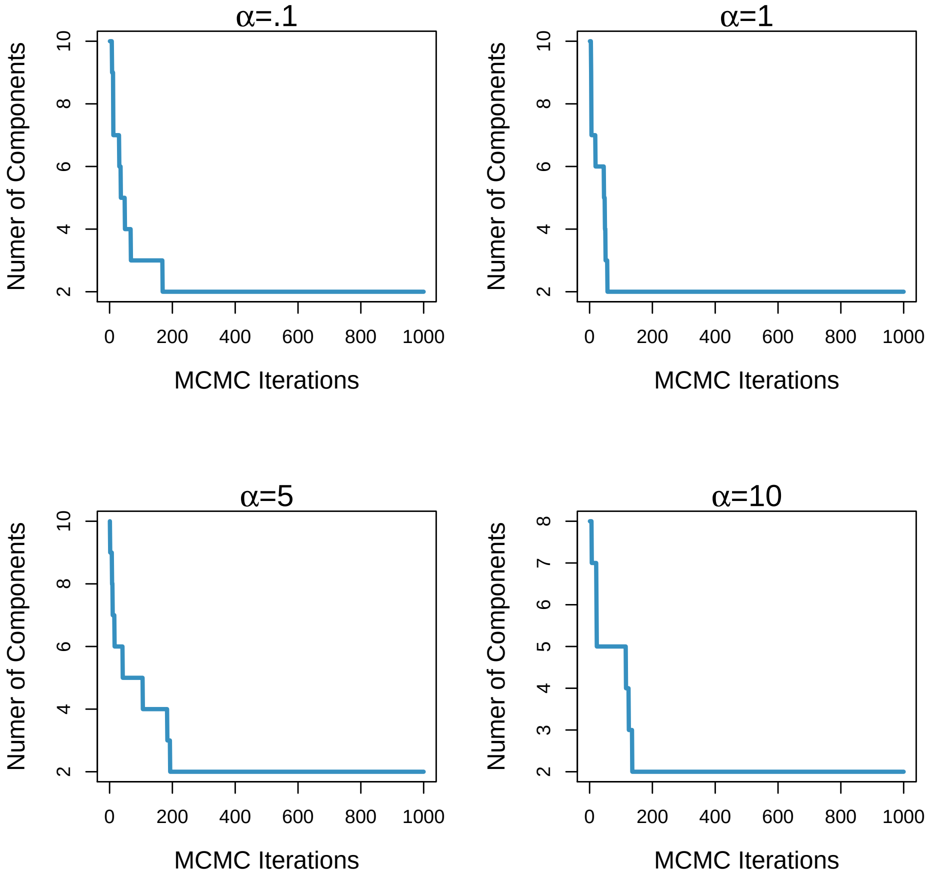

Figure 4.4: Sensitivity of Dirichlet process mixture models to values of α. Data was simulated as drawn from a mixture of two Exponential distributions which differ in time-scale by 10. The result of model inference for several fixed values of α. It is clear that over this range of α, the effect on the inferred models is negligible as the two component mixture is correctly inferred in each case.

4.4 Results

I now demonstrate the use of nonparametric Bayesian inference with diverse estimation problems in the study of single molecule biophysics.

4.4.1 Single Ion Channel Dwell-Time Distributions

First, I demonstrate the use of infinite mixture models to analyze dwell-time distributions from single ion channel recordings. Shown in Figure 4.5 are example recordings from a single BK channel at different holding voltages and at different calcium concentrations. As the BK channel is gated by both voltage and calcium, we see that increases in holding voltage or in calcium concentration result in increased open probability of the channel. The BK channel has been studied extensively by many groups and detailed mechanistic models have been put forth to explain the effects of voltage and calcium on channel gating (Cox et al., 1997; Rothberg and Magleby, 2000; Horrigan and Aldrich, 2002; Rosales and Varanda, 2009). This detailed understanding of BK channel gating provides an excellent testbed for the use of the these novel analysis methods.

As a first step to analyzing single ion channel recordings, we can deconstruct the time series into sojourns within closed states and open states (Colquhoun and Hawkes, 1981). To do this, I first de-noise, or idealize, the single channel data by classifying each datapoint as corresponding to either *closed* or *open*. The simplest method for this would be choosing a threshold of halfway between the average open and closed current levels and then classify-

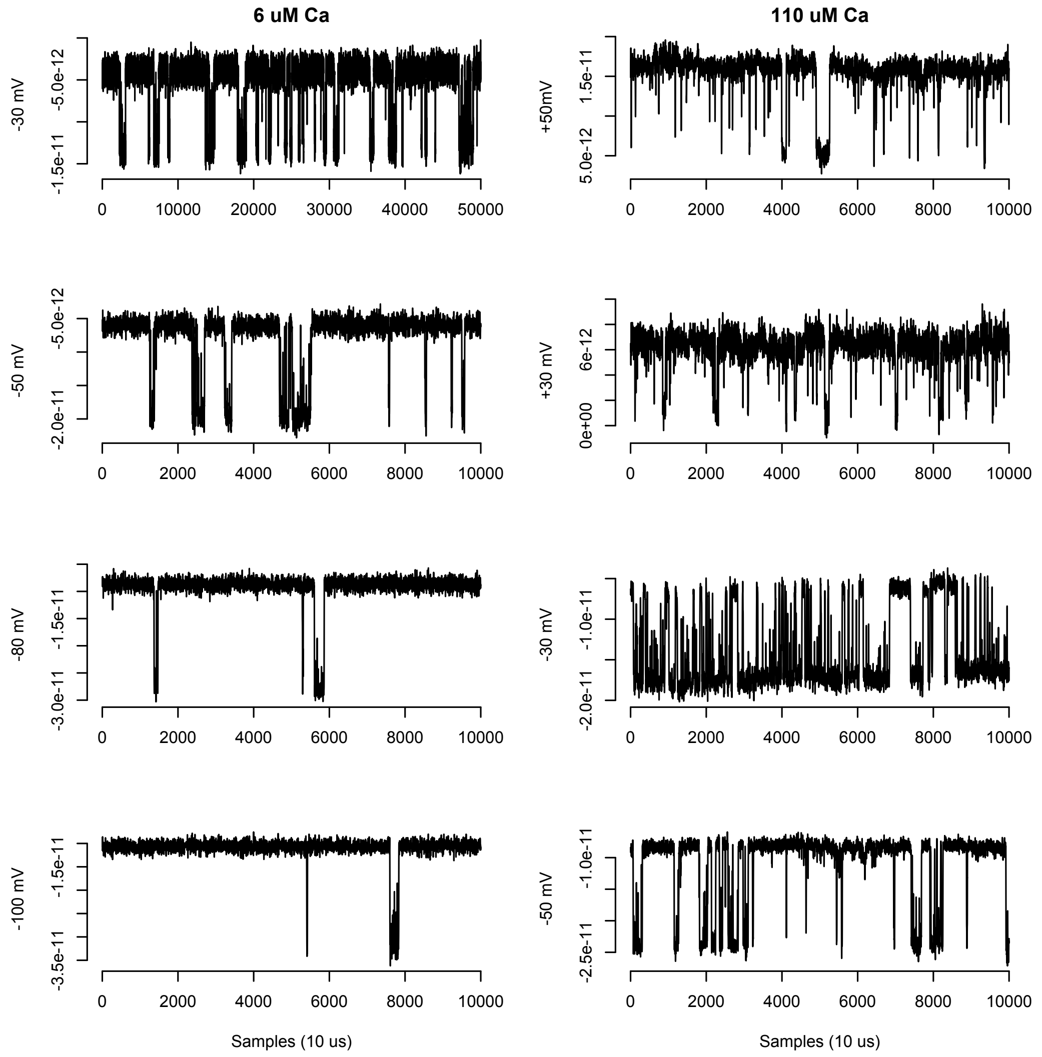

Figure 4.5: Example data from a single BK channel at various holding voltages and calcium concentrations. Patch currents reported in Amperes. Consistent with previous BK work, increases in calcium or transmembrane voltage will increase the open probability of the channel.

ing each data point relative to this threshold (Colquhoun and Sigworth, 1983). This simple method works fairly well, though one must be wary of artifacts such as threshold-crossing due to poor SNR and correcting for missed events (Colquhoun and Sigworth, 1983). I prefer an alternative approach, where we treat the time series as a two-state hidden Markov model. Here, the open and closed states each correspond to different levels of current obscured by noise, each with different variability. Notice that the threshold method would yield very similar results to any model with a symmetric noise distribution, but makes the assumption that the current variance is the same for both open and closed states. I prefer not to make that assumption and so model closed and open states corresponding to Normal distributions each with distinct mean and variance. Using the Gibbs sampling approach described in the Theory section, I utilize a latent indicator variable $s_1, ..., s_N$ to denote the state assignment for each data point. Thus, after MCMC inference, the indicator variables $s_1, ..., s_N$ yield the idealized trajectory through the hidden states. This Bayesian approach to idealization of ion channel records has been used previously and was thoroughly compared to previous methods (Rosales, 2004; Siekmann et al., 2011), so I omit such a discussion here. Figure 4.6 shows an example of this method. The data points are overlaid with colors corresponding to which conductance state (closed or open) each point was likely drawn from. With an idealized trace, we simply count how many consecutive samples are spent in a state before transitioning to the other state: this is a dwell time in one of the states. Decomposing the whole recording in this way yields a distribution of dwell-time events in the open state and in the closed state.

The theory of Markov processes indicates that the ensemble of dwell times should be exponentially distributed, if there is truly only one closed state and one open state. Figure 4.7 shows example dwell times from a recording of BK at 6 uM Ca and -30 mV. The ensemble of open times indeed appears to be exponentially distributed with a timescale of about 1ms. For the closing times, however, things are somewhat more complex. Note that while these data generally seem to be exponentially distributed, there is a large fraction of these events that occur within the first histogram bin. In fact, this is a cluster of events that happen on a faster timescale from the rest. Though a single channel time series implies the presence of only two conductance states, this dwell-time distribution indicates that there exist multiple states which appear as closed yet which have measurably distinct dynamics. Given that we have measured a set of dwell-times, interpretation of this data is simply a matter of fitting to a (potentially) multi-component mixture of exponential distributions. If we can decide how many components are in the data, then many methods might be used for estimating the parameters of a finite mixture model (Colquhoun and Hawkes, 1981; Sigworth and Sine, 1987).

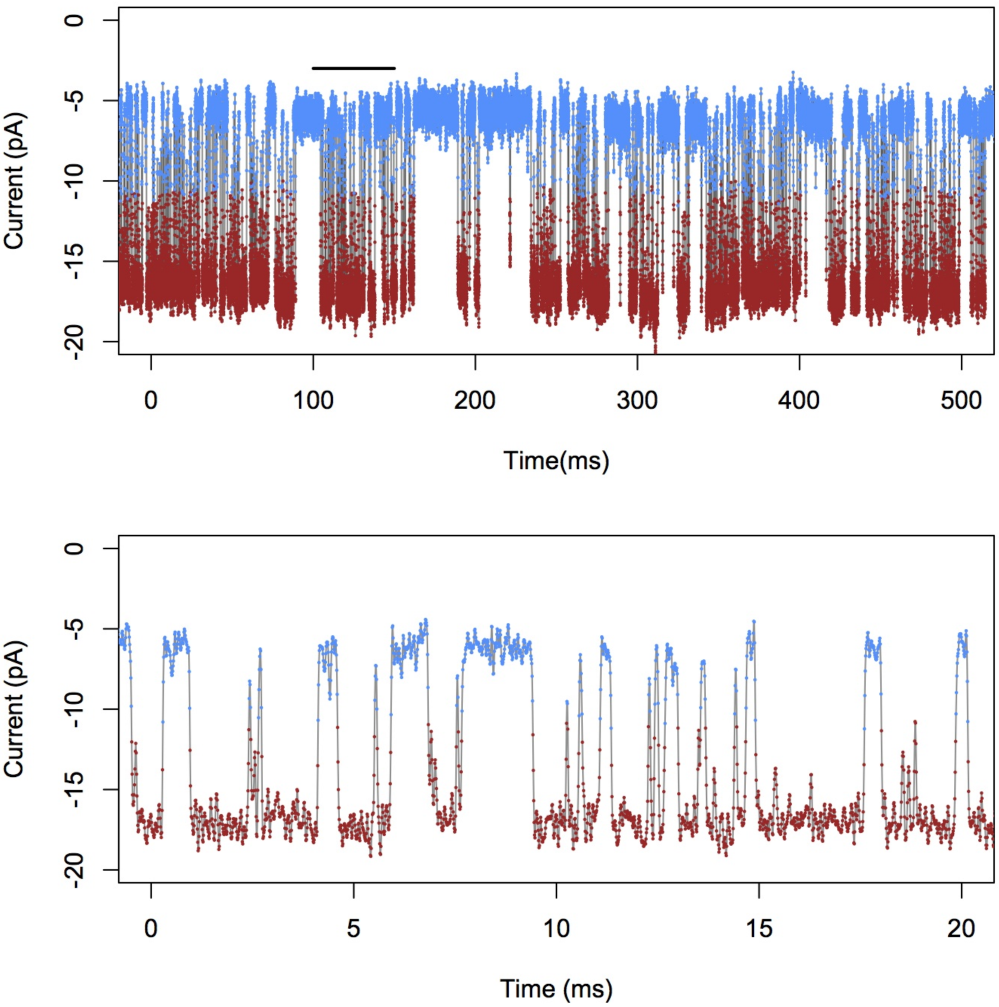

Figure 4.6: Using a two-state hidden Markov model to idealize single channel recordings. The time series is assumed to be drawn from a two-state Markov process where each state has a distinct emission distribution characterized by a Normal distribution with different means and variances. The model is fit using Gibbs sampling (see Theory) and the idealized trace (the hidden states) is shown as colors. Segments of the time series are shown at two different time scales.

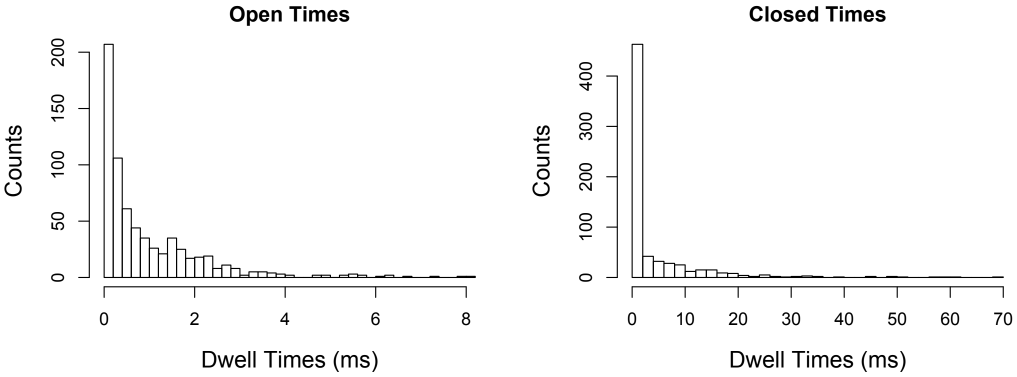

Figure 4.7: Dwell-time distributions. Example open- and closed-time distributions from a BK channel in 6 μM calcium held at -30 mV. Note that the histogram of closed-times indicates a large number of events occurring within the first bin. This indicates that there are closing events which occur on measurably distinct time scales. Hence, what appears to be simply open and closed states within the time series is actually indicative of many hidden states which have different dynamics. We can interpret dwell-time histograms as mixture of distinct Exponential components in order to estimate the properties of each hidden state.

Much effort has been put into data transformations and other methods for deciphering how many components exist in single channel dwell-time distributions (Sigworth and Sine, 1987; Landowne et al., 2013). Discovering the number of components within such data is an ideal use for Dirichlet process mixture models. As described in the Theory section, we imagine that the data y_i are drawn from infinite number of exponential components by using a Dirichlet process prior on the mixture weights,

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) && (4.42) \\ y_i &\sim \int p(y_i|\theta)G(d\theta) && (4.43) \\ &= \sum_{i=1}^{\infty} w_j e^{-(\theta_j/y)}. && (4.44) \end{aligned}$$

By using this infinite model to fit our finite data, we are able to discover the number of components in the data, instead of assuming it. In the Theory section, I demonstrated that this model could indeed discover the number of components in simulated data drawn from mixtures of exponential distributions and it could also provide accurate estimates of the relevant parameters and their uncertainties (see Figure 4.2). This method can be applied to dwell-times from BK channel recordings at various holding voltages. Figure 4.8 shows dwell-time distributions from 5 seconds of data from a BK channel in 6 μM calcium held at several voltages. These dwell-time distributions have been analyzed using an infinite mixture of Exponential distributions so that we can discover the number of components in the data, instead of pre-supposing it or fitting many different models sequentially. In Figure 4.8, the dwell-times are visualized as histograms and are also shown in the rug plots below the histograms. The color of the data points corresponds to the component from which they were likely drawn and the probability density of each component is shown atop the histogram. Finally, the total probability density from all components is shown as the grey trace, which overlays well with the observed histograms. We are able to extract from these data the number of hidden components and that these results are consistent with what is previously known about the BK channel. For example, we see that with the three increasing holding voltages, the infinite mixture model indicates the emergence of measurably distinct open states. Additionally, we have a rigorous estimate of the time-scale parameter for each component, and can see that the alterations in mean dwell time (as a function of voltage) are consistent with previous findings (Cox et al., 1997). This task of determining the number of significant components in dwell-time distributions is easily accomplished using a Dirichlet process mixture model (see Discussion for comparison with previous methods).

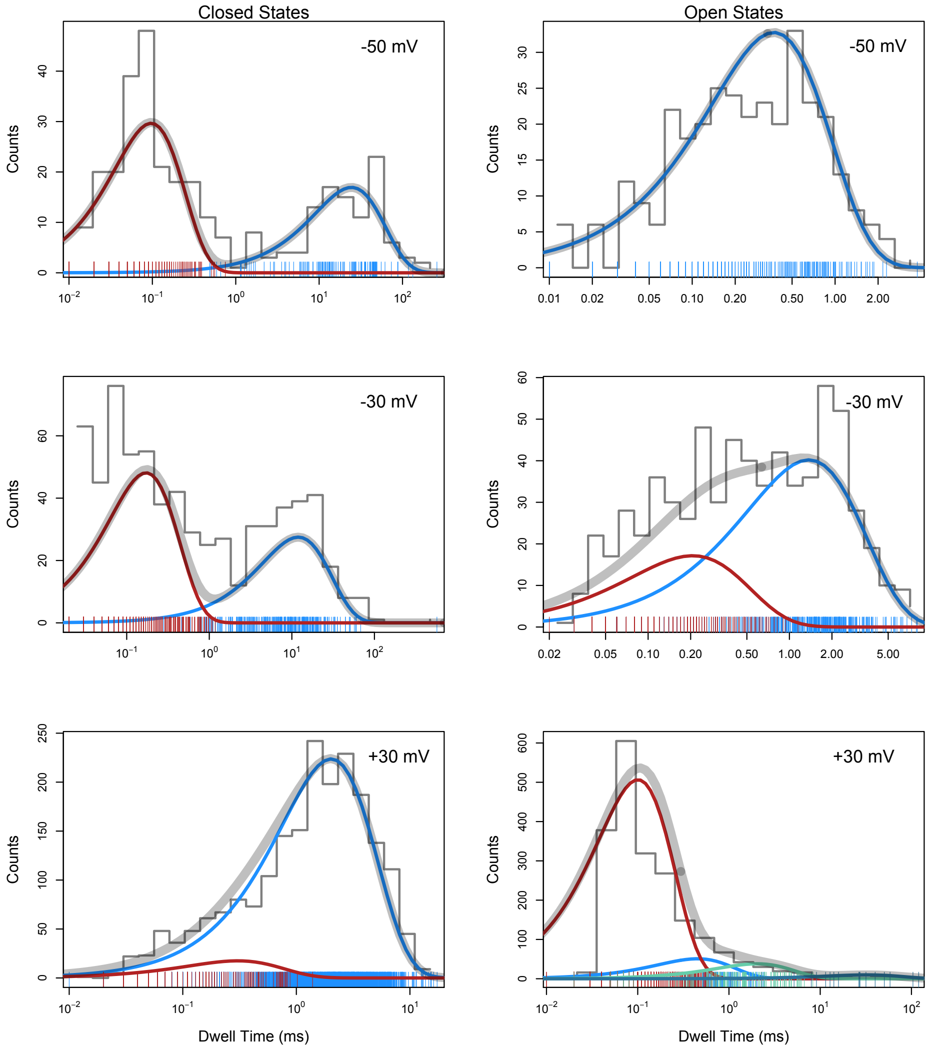

Figure 4.8: Dwell-Time Distributions and Infinite Mixture Models. These dwell-times, plotted logarithmically for visualization, are from 5 seconds of a BK channel at 6 μM calcium and various holding voltages. Distributions of dwell-time are analyzed with an infinite Exponential mixture model in order to discover how many components are in the data. The color of the data points corresponds to which component they belong to and the probability density of each component is shown atop the histogram. Finally, the total probability density from all components is shown as the gray trace. Algorithm parameters: $\alpha = 1$.

4.4.2 Application of iHMM to Single Molecule Time Series

Hidden Markov models (HMM) have enjoyed vast application in many areas of science and engineering due to their flexibility and predictive ability (Rabiner, 1989). For stochastic time series arising from single molecule measurements, we might imagine that the observations y_t are Normally distributed random variables and that each hidden state corresponds to a normal distribution with a different mean and precision such that $y_t \sim N(\theta_i, \frac{1}{\tau_i})$. As described in the Theory section, a nonparametric Bayesian extension of this HMM framework is the hierarchical Dirichlet process hidden Markov model (Beal et al., 2002; Teh et al., 2006). Using this model, we do not fix the number of hidden states before data analysis, but instead we can learn the number of hidden components within the data. An example use of this model is shown in Figure 4.9. The top row shows an electrophysiological recording from a patch which contains a unknown number of ion channels. The holding voltage is negative, so downward deflections of current indicate events of ion channel opening. From this multi-channel patch, we might want to estimate the number of channels in the patch and the average open probability. When different numbers of channels are open at different times, we observe this as distinct levels of current, obscured by electrical noise. Thus, we can use an infinite hidden Markov model (iHMM) approach to learn how many distinct current levels exist in the time series and the number of channel openings seen. Figure 4.9 (bottom) shows the result of iHMM modeling, with each data point colored corresponding to which hidden state it is likely drawn from. It is clear that we are able to correctly detect the number of distinct levels of current and infer the number of open channels seen in this patch. In this particular case, the signal-to-noise of the recording is quite high and we could perform this task by eye fairly easily, but it serves as a general demonstration of the kinds of data are well suited for the iHMM.

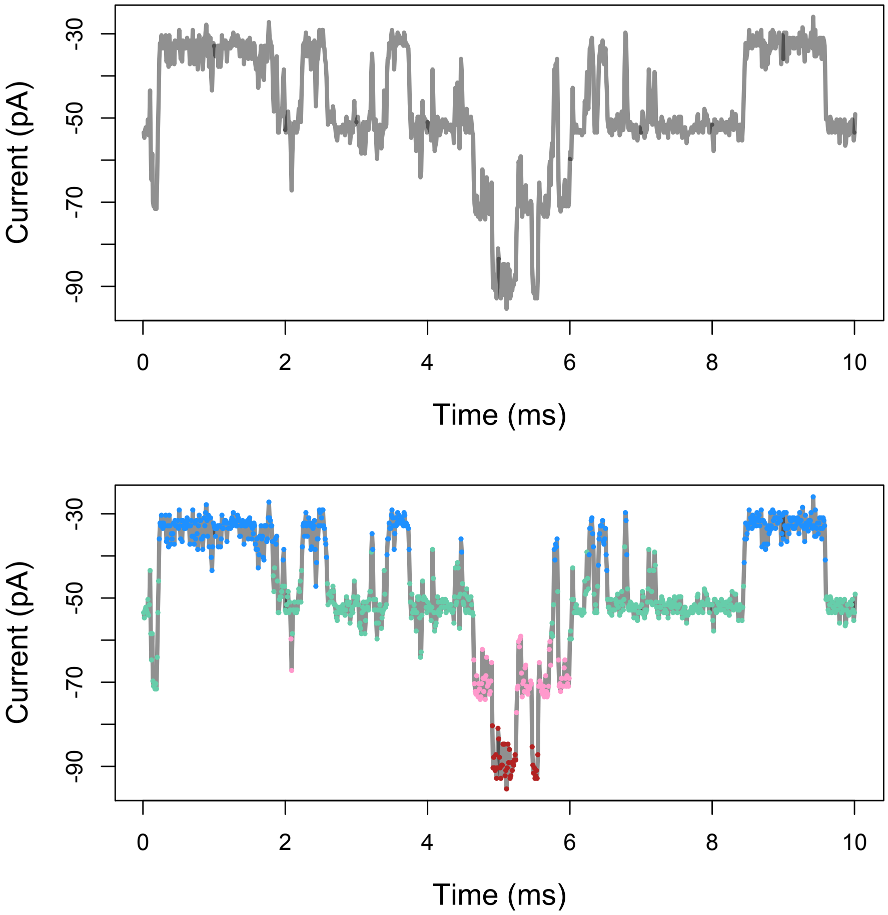

Figure 4.9: Application of iHMM to multichannel patch recording. (Top) An example recording of a patch that contains multiple ion channels. Hold voltage is negative, so downward deflections of current are indicative of channel opening. (Bottom) Colors indicate which hidden state each datapoint is assigned to. iHMM is able to correctly determine the number of channels in the patch. Algorithm parameters: $\alpha = 1, \gamma = 1$.

As a more challenging application, I use the iHMM to de-noise single molecule FRET traces and decipher distinct conformational states and transitions. Figure 4.10 shows such single molecule FRET traces recorded from the agonist-binding domain of the NMDA receptor (see Methods). In the traces shown (left column), we can see that the FRET efficiency indicates the molecules tend to reside within distinct conformational states for tens of milliseconds before transitioning to other states. However, the noise in this data makes it difficult to tell when these transitions occur and, more importantly, how many conformational states are observed within each trace. We can use the iHMM to analyze these traces in order to detect the presence of significant conformational states. Figure 4.10 shows the data overlaid with colors according to which hidden state each data point was likely drawn from. The iHMM is able to decipher distinct conformational states based on both the properties of the emission distribution (mean and variance) as well as the dynamics of the states. Even if a state is visited extremely rarely (such as in the top trace), we are able to confidently assert the existence of distinct conformational states. The right column of Figure 4.10 shows the posterior distribution over number of states for each trace. This posterior probability provides a simple way to quantify confidence in an interpretation of the number of states and we can use the posterior maximum to inform us about the most probable number of distinct states in the data. An interesting extension of this model would be to combine an ensemble of different traces into a hierarchal model (van de Meent et al., 2013). In such a model, we imagine that each trace provides a brief snapshot of some underlying hidden distribution from which all the traces are drawn. Then the traces, taken in aggregate, provide information about the total conformational space and transition dynamics. Future work remains to be done in this area.

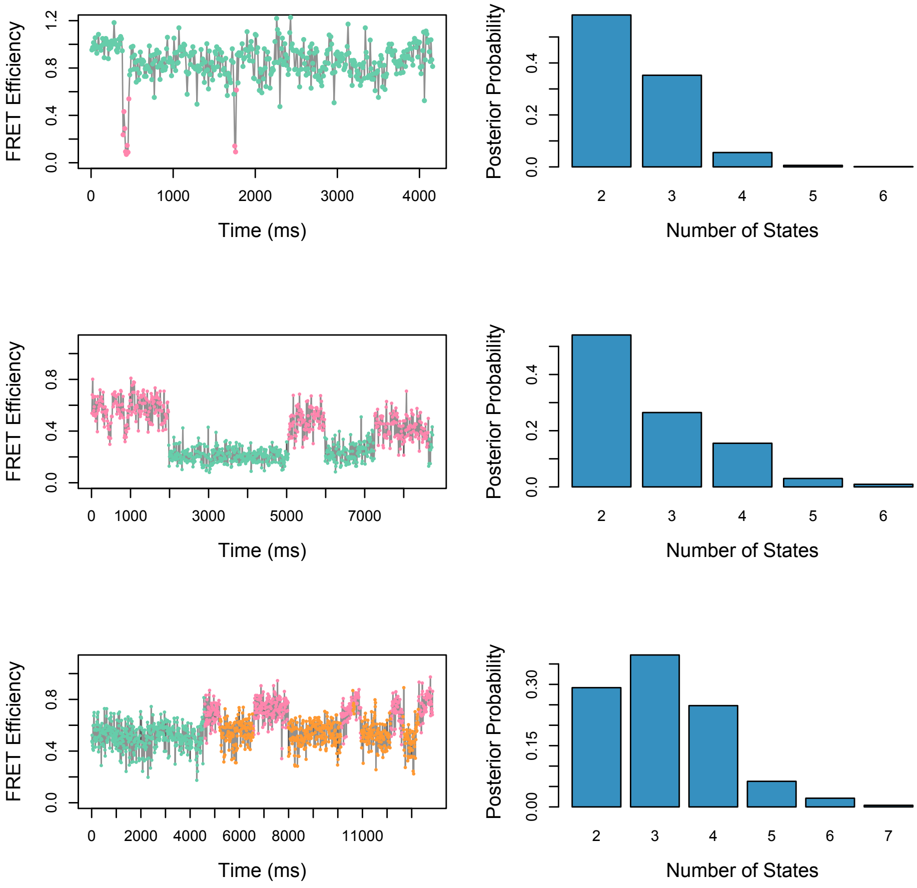

Figure 4.10: Application of iHMM to single molecule FRET. (Left column) Example traces of FRET efficiency over time. Sudden conformational changes are evident, but it it difficult to know the number of states and precise moment of state changes in these noisy traces. Colors indicate which hidden state each data point is assigned to. (Right column) Posterior distributions over number of hidden states inferred for each trace. The iHMM is able to decipher the number of the number of conformational states represented in these noisy time series. Algorithm parameters: $\alpha = 1, \gamma = 1$.

As a final application, I turn to single molecule photobleaching. In this setting, we observe photon counts over time and are interested in detecting photobleaching events which reveal themselves as sudden decreases in photon intensity. We are particularly interested in counting the number of photobleaching events in a data trace. This setting is well suited for the iHMM since we want to detect transitions between an unknown number of states (corresponding to bleaching events). Figure 4.11 (left column) shows example traces. We can see that photobleaching events are apparent, but in regimes of low signal-to-noise, it might be quite difficult to tell by eye when bleaching events occur. After analysis with the iHMM, the data points are colored corresponding to the hidden state to which they were assigned. It is clear that the iHMM is an excellent tool for this task. Even in settings where photobleaching events are very difficult to detect by eye (bottom), the iHMM is able to determine the number of transitions in the data. The right column shows posterior distributions over the number of hidden states in the data, which provides a natural quantification of confidence when interpreting this data. Using the iHMM provides a rigorous and unbiased method to analyze these traces.

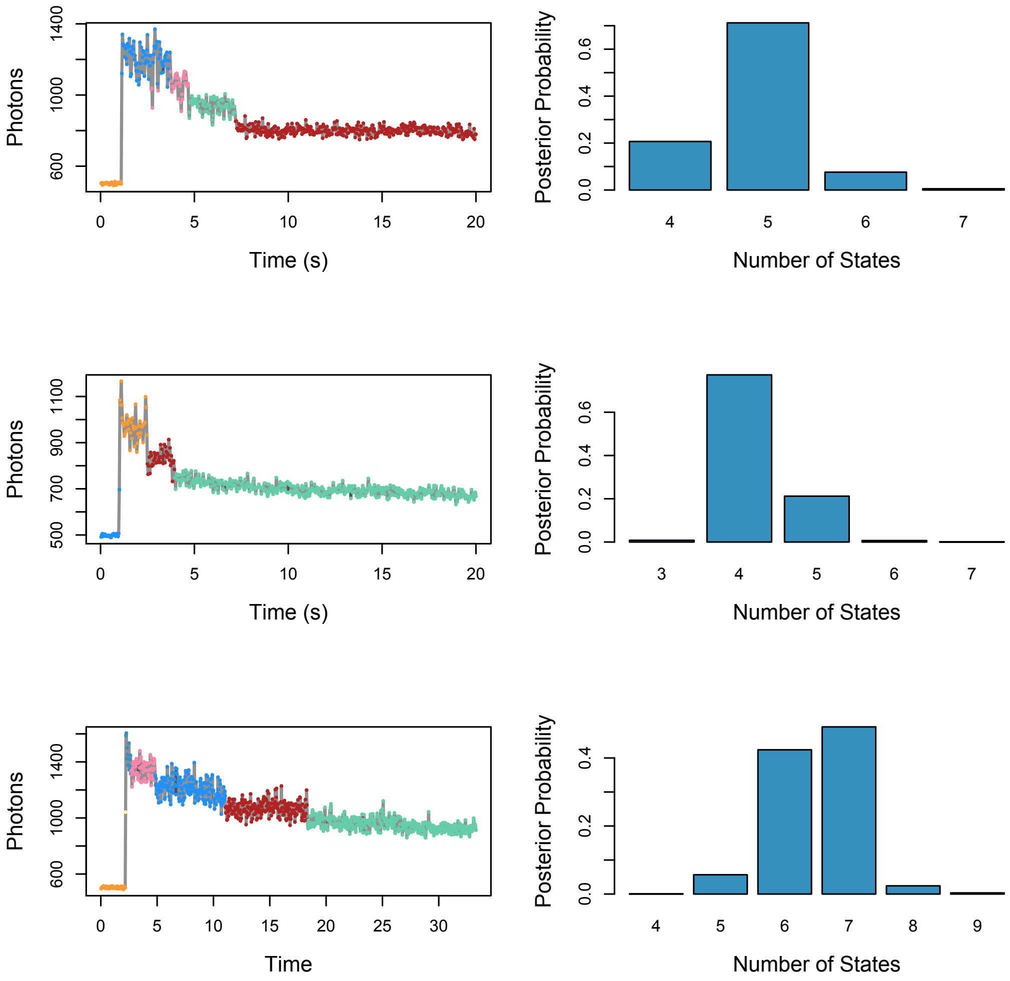

Figure 4.11: Application of iHMM to single molecule photobleaching. (Left column) Example traces of photon counts over time (sampling rate 30 Hz). Sudden photobleaching events are evident, but it it difficult to know the number of bleaching steps in the presence of noise. Colors indicate which hidden state each data point is assigned to. (Right column) Posterior distributions over number of hidden states inferred for each trace. The iHMM is able to decipher the number of the number of bleaching events and also provides a quantification of confidence. Algorithm parameters: $\alpha = 1, \gamma = 1$.

4.4.3 Application of iAMM to Single Ion Channel Recordings

Next, I demonstrate the use of the infinite aggregated Markov model to analyze single ion channel recordings. I previously analyzed BK single channel data by deconstructing the time series into dwell-times and fitting Exponential mixture models. This approach throws away much information and a preferable method is to model each time point in Markov-type model (Qin et al., 1997). In the Theory section, I introduced the infinite aggregated Markov model (iAMM), where we assume a degeneracy such that multiple hidden states share an emission distribution. I demonstrated that the iAMM (more precisely, the *sticky*-iAMM, see Theory section) can be used to *learn* the number of hidden states from an AMM time series (Figure 4.3). I now apply this to single BK data where we see stochastic transitions between open and closed states of the channel, but suspect there exist more than two hidden states. Using the iAMM , we can learn the presence of open and closed states in the time series, instead of assuming this beforehand. For Figure 4.12, I have used 1 second of a recording of a single BK channel in 110 μM calcium held at +30mV. The top row of Figure 4.12 visualizes the posterior distribution over the number of hidden states and we see that we infer the presence of four hidden states. Figure 4.12 (middle) shows the data trace that was analyzed and the colors correspond to the hidden state form which each data point was likely drawn. This analysis reveals one open state and three closed states. We can see that there is a fast closed state (green) and a measurably slower closed state (red). Additionally, there is an extremely slow closed state (pink), of which we have only one ”observation”, but are easily able to infer its existence due to its distinct temporal dynamics. The bottom trace is the same data at an expanded time scale. Encouragingly, we are able to detect the presence of distinct hidden states based solely on their dynamical differences in single ion channel recordings.

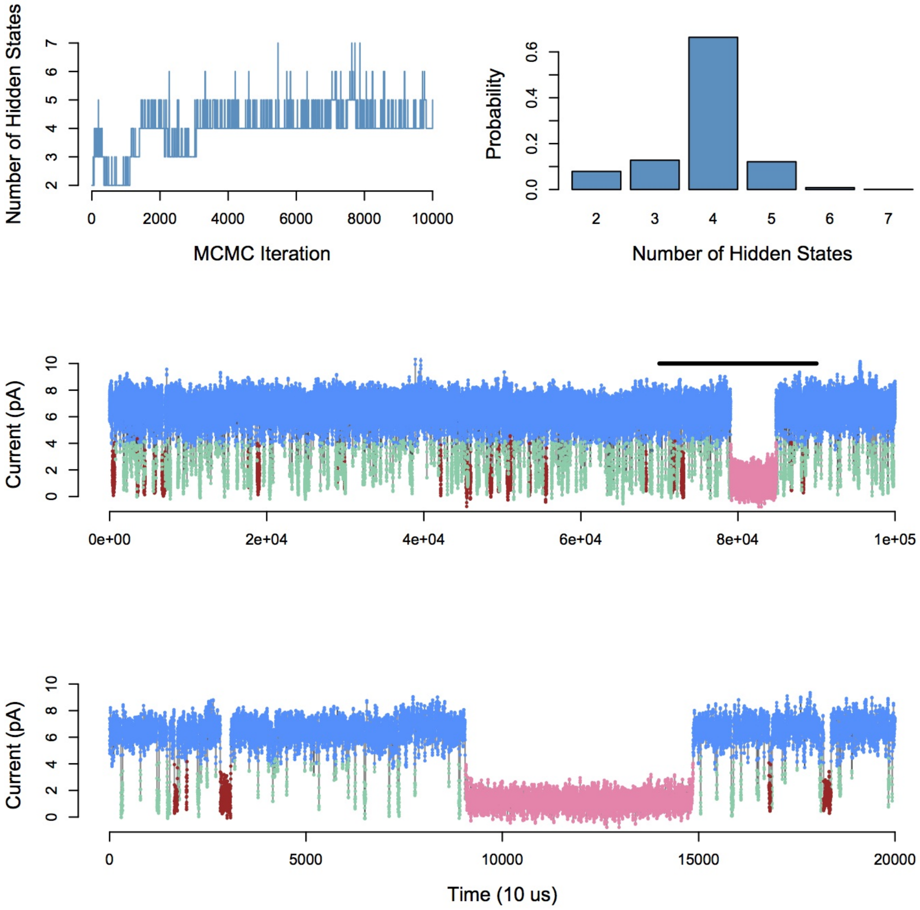

Figure 4.12: Application of iAMM to BK data. (Top row) The posterior distribution of number of hidden states indicates that this data has four hidden states. (Middle) Data trace with each data point labelled according to the hidden state from it was likely drawn. The iAMM finds one open state (blue), and three closed states. For the closed states, the fastest time-scale state (green) is different enough from a slower one (red) that we are able to identify them as distinct. Additionally, an extremely slow closed state (pink) is identified. (Bottom) Same data at an expanded scale. Algorithm parameters: $\alpha = 1, \gamma = 1, \kappa = 100$.

Before continuing, I describe a very general barrier to the analysis of single channel recordings which the iAMM is still unable to surpass with the present dataset. (Kienker, 1989) noted that with aggregated Markov models, the space of potential mechanisms that can adequately fit any given equilibrium data is non-identifiable. In particular, it was shown that the transition matrix π of any given AMM exists amongst a (possibly infinite) equivalence class of other $\tilde{\pi}$ which would all produce identical data. Not only might it be impossible to derive a unique estimate of π from data, but the members of such an equivalence span a continuous range of $\tilde{\pi}$, including members with entirely different connectivities between the hidden states. Hence, the problem of model selection is exacerbated, since many different models (different connectivities) can be transformed into one another and would all fit the data equally well. In fact, the only way to circumvent non-identifiability in aggregated Markov models is to pre-suppose a particular connectivity. Often, such a constraint on the connectivity between the states allows typical inference methods, such as maximum likelihood (Qin et al., 1997) or Gibbs sampling (Rosales, 2004), to yield a unique estimate of π conditioned on a particular model. However, since my goal has been to avoid the pre-specification of models, the iAMM approach will inevitably suffer from this model non-identifiability when fitting equilibrium time series. It is likely that this non-identifiability can be overcome by using non-stationary methods (Kienker, 1989; Millonas and Hanck, 1998; Milescu et al., 2005) and future work remains to be done in this area.

Despite this limitation, the iAMM approach can still be used to gain qualitative insights and to test the algorithm against what is previously known about the BK channel. In order to visualize the results, I cast the inferred state topology into a canonical form which is representative of, and unique to, a particular equivalence class. Several such canonical forms have been proposed including uncoupled form (Kienker, 1989), manifest interconductance rank form (Bruno et al, 2005), reduced dimensions form (Flomenbom and Silbey, 2006), and maximum entropy form (Li and Komatsuzaki, 2013). Since I am using canonical forms solely for visualization, and not to estimate the resulting transition rates, I use the Kienker uncoupled form due to its simplicity. Here, the connectivity is shown in the simplest form where none of the states of the same aggregate are connected to each other. That is, open states are only connected to closed states, and vice versa.

Figure 4.13 shows the result of using the iAMM to analyze several BK recordings at 6 μM and 110 μM calcium. The data visualized here represents a small fraction of the full trace used for model inference which was 1 second of data (10^5 samples) in each case. At left, the data traces are shown with data points colored corresponding to which hidden state they are likely drawn from. At right, the model inferred from each trace is shown in Kienker uncoupled form. Again, while the state topology shown here is but one of many which could explain the data with high posterior probability, the visualization is used here to convey the general complexity of the gating mechanism which generated each trace. At very low holding voltage (-100 mV), open probability is very low, but also the available state space explored by the channel is as simple as possible, with one open state and one closed state. As the holding voltage is increased, not only does open probability increase, but also we detect the presence of more open and closed states. The increase in voltage affects channel function not only by shifting the open probability, but by allowing the channel to access a more complex state space. As holding voltage is increased further, and open probability begins to saturate at a high value, the complexity of channel gating decreases, as the channel accesses fewer conformational states. The last trace in Figure 4.13 is at +30mV and 110 μM calcium, which is in an extreme corner of BK's activation range. The open probability is very high and in this extreme range the complexity is decreased, as the channel mostly occupies a single open state with infrequent sojourns to just two closed states. Consistent with what is known about the BK channel (Rothberg and Magleby, 2000; Talukder and Aldrich, 2000; Horrigan and Aldrich, 2002), we see that in the extreme ranges of voltage and calcium, characterized by either very high or very low open probability, the channel gating landscape is the least complex. Conversely, in the middle of the activation range, the gating scheme is most complex, with the channel accessing a diversity of open and closed states. Using a nonparametric Bayesian approach, we were able to recover this fundamental principle of channel gating, by discovering structure hidden within these time series.

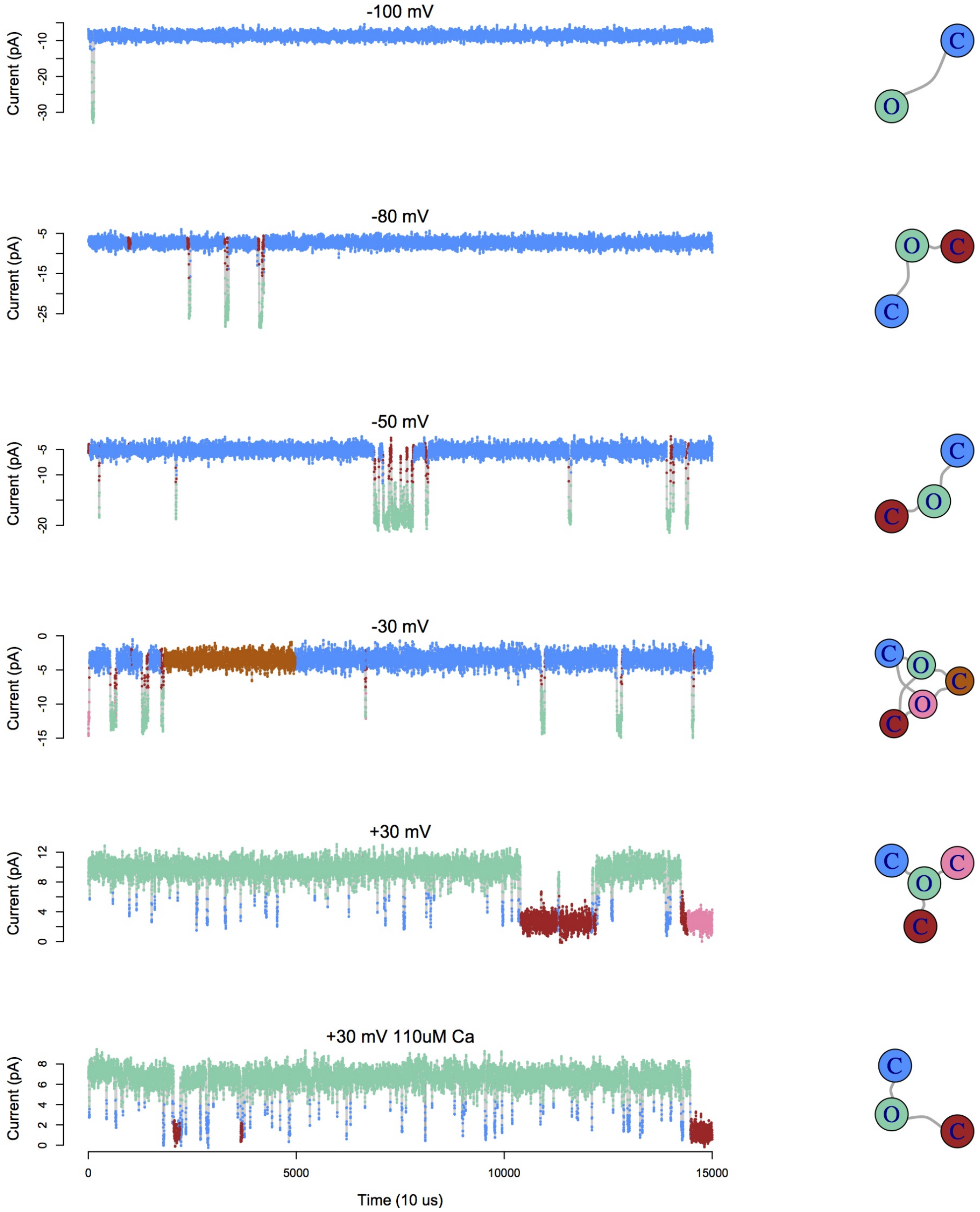

Figure 4.13: Recordings from a BK channel at multiple holding voltages and calcium concentrations analyzed using the *sticky*-iAMM. Data points are colored corresponding to the hidden state from which they were drawn in the inferred model. At right, the inferred model for each trace, visualized in Kienker uncoupled form. Algorithm parameters: $\alpha = 1, \gamma = 1, \kappa = 100$.

4.5 Discussion

Here I have introduced the use of nonparametric Bayesian inference for the study of single molecule biophysics. These methods rely on the properties of the Dirichlet process in order to employ an infinite dimensional probability distribution. When used to model finite data, this infinite model effectively allows us to discover structure in data instead of assuming it beforehand. The power and flexibility of these methods was demonstrated with diverse applications in single molecule biophysics.

I demonstrated a basic use of nonparameteric Bayesian inference by using a Dirichlet process mixture model to analyze dwell-times from single ion channel recordings. Using this infinite mixture model, it is possible to discover how many hidden clusters lie within the data and in this way, the number of hidden states could be learned, instead of assumed. For the case of ion channel dwell-times, much emphasis has been placed on optimal methods for analyzing such data (Colquhoun and Hawkes, 1981; Sigworth and Sine, 1987). Recently, (Landowne et al, 2013) described a method to fit dwell-time data without knowing the number of components. This is similar in goal to the infinite mixture model described here. Their approach, grossly paraphrased, consists of: beginning with a number of components which is very large (they use 20), iteratively using maximum likelihood to optimize the timescale and weight parameters of each component, removing clusters which are deemed to be too similar in timescale (they chose 2%) or too small in weight (they chose 10^{-5}), continuing this process of removing clusters until the log-likelihood is no longer improved. They demonstrate that their approach works very well to correctly identify the number of components in simulated data as well as BK channel data. Their approach, while convincingly demonstrated and validated, is not based on a rigorously defined mixture model, but instead consists of iterative hypothesis testing, ad hoc thresholds, and parameter optimization until the fit to data no longer improves. In contrast to this, the Dirichlet process mixture model is rigorously defined over an infinite set of mixture components, however, the properties of the Dirichlet process guarantee a clustering of the data. With channel data, a small number of distinct clusters is detected with high posterior probability. By sampling the space of all mixture models, we calculate the posterior distribution over the number of clusters in the data and can quantify our confidence in an interpretation of the data. Further, by sampling the full posterior (as opposed to simply seeking a maximum likelihood estimate), we can address parameter non-identifiability, a pitfall which is sure to be problematic for exponential mixtures and small sample sizes.

In addition to channel data, (Landowne et al., 2013) test out their methods with classic datasets which have been deemed to be extremely challenging. They show that their method does very well in all cases to correctly detect the number of components. For comparison and validation, Figure 4.14 shows the result of using an infinite mixture model to analyze each of these data sets. Boliden 3 corresponds to a mixture of four exponentials where each component has higher timescale parameter and larger weight. Boliden 4 is a mixture of four exponentials where one of the components has smaller weight than the adjacent components. Figure 4.14 shows the result of using the infinite mixture model to analyze $N = 10000$ data points drawn from each distribution, using the parameter values reported in Tables 2 and 3 of (Landowne et al., 2013). It is clear that we infer, with high posterior probability, the correct number of components in each case. Importantly, we can detect the presence of these components using only 10^4 data points, which is a thousand-fold smaller sample size than the 10^7 samples used by (Landowne et al., 2013). While it is clear that the approach of (Landowne et al., 2013) works very well with large datasets, and is almost certainly faster than an MCMC-based approach, they discuss the limitations of their hypothesis-testing based approach when faced with inadequate sample size. In this small-sample regime, the Bayesian approach presented here will be much better able to detect significant components in the data.

A generalization of mixture models might be one where we do not assume each datapoint is drawn independently from the underlying distributions, but instead we assume there is dependency between successive data points which is governed by some Markov process. Such a hidden Markov model is a popular tool for modeling stochastic time series and I showed how the nonparametric Bayesian extension, the hierarchical Dirichlet process hidden Markov model, can be successfully applied to single molecule time series. Using an iHMM to analyze multi-channel patch recordings allows us to estimate the number of channels and open probability in noisy electrophysiological data. Additionally, I used the same model to analyze data from the increas-

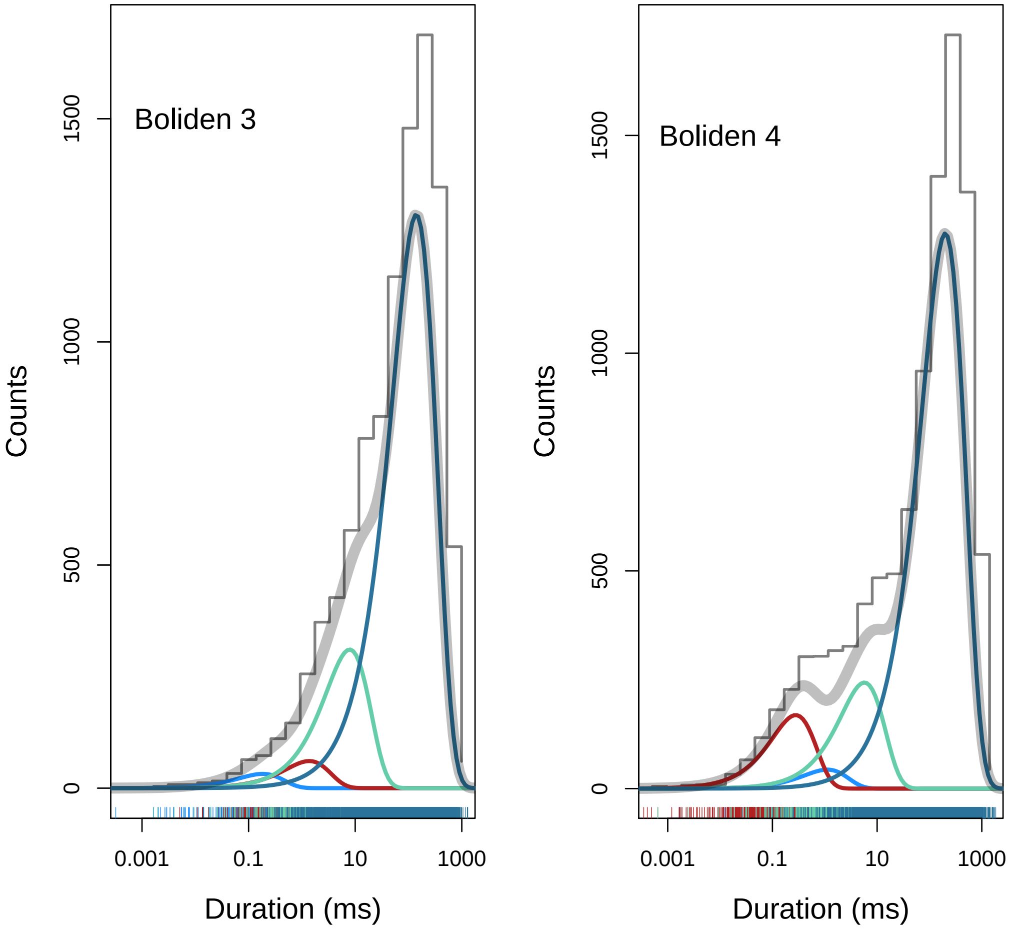

Figure 4.14: Demonstration of the infinite Exponential mixture model with data sets discussed in (Landowne et al., 2013). Boliden 3 corresponds to a mixture of four exponentials where each component has higher timescale parameter and larger weight. Boliden 4 is a mixture of four exponentials where one of the components has smaller weight than the adjacent components. Using the parameter values reported in (Landowne et al., 2013), 10^4 data points are drawn from each mixture and analyzed using the infinite model. We are able to correctly recover the number of components and the relevant parameters.

ingly popular method of single molecule FRET. In this case, we are interested in detecting distinct conformational states, as manifest in the noisy FRET efficiency signal. We can use the iHMM to analyze these traces in a typical hidden Markov approach, but without assuming the number of distinct states or their properties. Finally, I showed that single molecule photobleaching traces can be analyzed with the iHMM in order to detect bleaching steps. Especially in cases of poor signal-to-noise, the iHMM provides a principled method to analyze such data. Generally, using the infinite hidden Markov model provides a rigorous and unbiased method to interpret these time series.

I showed that a special case of the iHMM, the infinite aggregated Markov model, could be used to analyze single ion channel recordings in order to detect the existence of hidden conformational states. I showed that this approach can be used to infer the presence of distinct open and closed states which differ only in their dynamics. Further, when this approach is applied to BK channel recordings at multiple calcium concentrations and holding voltages, the inferred gating schemes recapitulate basic principles regarding the complexity of BK channel gating. However, with the equilibrium single channel traces, we are still limited by non-identifiability and cannot infer a unique and reliable estimate of the connectivity between these hidden states. Previous authors have shown the benefits of globally analyzing large data sets in aggregate or of incorporating non-stationary stimulus protocols (Kienker, 1989; Millonas and Hanck, 1998; Milescu et al., 2005; Rosales and Veranda, 2009). I suspect that such a strategy, coupled with an iAMM type approach, may be all that is required to overcome the barrier of non-identifiability and be able to extract accurate and reliable models of ion channel gating from single molecule recordings. Future work remains to be done in this area.

The study of protein biophysics has been greatly aided by the emergence of single molecule experimental techniques, but developing rigorous and general tools for the analysis of such data remains an open challenge. I have described the use of nonparametric Bayesian inference, a powerful paradigm which has gained recent popularity in the statistics and machine learning communities and which has been applied successfully to many difficult problems in science and engineering. These tools allow us to side-step the problems of model selection and user bias and instead allow us to discover significant structure in data, instead of assuming it beforehand. This framework was demonstrated with diverse settings in single molecule biophysics, with models including nonparametric mixture models, hidden Markov models, and aggregated Markov models and data sets including electrophysiology, single molecule FRET, and single molecule photobleaching. This paradigm provides a powerful basis to enhance the study of protein biophysics.

Appendices

Appendix A

A Primer on Bayesian Inference for Biophysical Systems

A.1 Introduction

The proper interpretation and analysis of experimental data is vital in the endeavor to understand natural phenomena. Here, I describe the use of Bayesian inference, a statistical paradigm which has gained popularity in many fields including astrophysics (Loredo, 1990), systems biology (Klinke, 2009), econometrics (Geweke, 1989), and many others. However, the adoption of Bayesian methods has been relatively slower in the study of protein biophysics, a field which relies primarily on more classical techniques. It is not my intention here to argue the merits of Bayesian methods over others, as this has been discussed elsewhere (Siekmann et al., 2012; Calderhead et al., 2013; Hines et al., 2014). Instead, my aim is to provide an accessible introduction and tutorial on the use of these methods with a focus on problems which should be familiar to the biophysicist.

A.2 Bayesian Inference

In Bayesian inference, the primary goal is to compute the *posterior distribution.* This is a probability distribution over the parameter space which quantifies how probable it is that a particular value of the parameter(s) gave rise to the observed data. This distribution provides not only an optimal point estimate (the *maxiumum a posteriori* or MAP estimate), but also a quantification of the whole parameter space, yielding a simple method for calculating confidence intervals. In this way, we consider the entire parameter space and ask which regions are most likely to be true, given the data we saw. For straightforward models, we can derive simple expressions for posterior distributions by using conjugate models. For more complex models, we can take advantage of computational methods that allow us to estimate posterior distributions of arbitrarily high dimension.

Consider that we treat not only the data y as random, but also treat the parameters of interest θ as random variables. From the definition of conditional probability, we can write

$$p(y|\theta) = \frac{p(y,\theta)}{p(\theta)} \tag{A.1}$$

$$p(y,\theta) = p(y|\theta)p(\theta). \tag{A.2}$$

We can also write the other conditional density,

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} \tag{A.3}$$

$$p(y,\theta) = p(\theta|y)p(y). \tag{A.4}$$

These are two expressions for the joint density $p(y, \theta)$, and we can equate them,

$$p(y|\theta)p(\theta) = p(\theta|y)p(y). \tag{A.5}$$

Rearranging this yields Bayes' rule,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{A.6}$$

By treating both the parameters and the data and random variables, a simple manipulation of conditional probabilities yields a general expression for $p(\theta|y)$, the posterior distribution of the parameters. The other components of Bayes rule are: $p(y|\theta)$, the likelihood of seeing the data given the parameters, $p(\theta)$, the prior distribution of the parameters, and $p(y)$, the marginal likelihood of the data. In practice, we generally only need to quantify the posterior distribution up to a constant of proportionality, so $p(y)$ is often ignored since it is independent of θ. This yields a more common form of Bayes' rule,

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{A.7}$$

Computing the posterior distribution is then simply a matter of deciding upon the likelihood and the prior distribution and combining them. I'll next show that if we put a little thought into finding prior distributions which are *conjugate* to the likelihood, then we can arrive at a simple expression for the posterior. Since we won't always be able to use a conjugate prior, I'll later discuss (at length) the powerful computational methods that allow us to calculate arbitrarily complicated posterior distributions.

A.3 Conjugate Models

I'll motivate our first foray into Bayesian modeling by taking as an example the experimental method of single molecule photobleaching (Ulbrich and Isacoff, 2007). This is a powerful method for determining the interaction and stoichiometries of protein complexes. The strategy consists of tagging a fluorescent probe to a protein subunit of interest and then imaging single molecules. After sufficient time, the fluorophores will photobleach, and by counting the number of photobleaching events, we get a direct readout of how many subunits are associated. However, there is a non-negligible probability that a fluorophore is already bleached before the measurement started. We'll quantify this probability of being pre-bleached as $1 - \theta$; that is, θ is the probability that a fluorophore bleaching event will be successfully detected. The result of this pre-bleaching is that a complex of n molecules might result in less than n bleaching events. Therefore, the ensemble of many such counts will be Binomially distributed such that the probability of seeing y bleaching steps when n are possible goes as,

$$p(y|\theta) = \frac{n!}{(n-y)!y!}\theta^{y}(1-\theta)^{n-y}. \tag{A.8}$$

As a simple inference problem, let's suppose that we want to estimate the pre-bleaching probability of an unknown fluorophore. To do this, we use a protein system that is well known so that we can assert that n is fixed to some known value. We perform a photobleaching experiment and gather N independent observations of bleaching counts and denote the total dataset as y_N. Our goal is then to estimate θ from y_N. Restating Bayes rules,

$$p(\theta|y_N) = p(y_N|\theta)p(\theta) \tag{A.9}$$

$$p(\theta|y_N) = \prod_{i=1}^{N} p(y_i|\theta)p(\theta). \tag{A.10}$$

Since we know the likelihood is a Binomial distribution, we can begin to fill in the components of Bayes' rule,

$$p(\theta|y_N) = \prod_{i=1}^{N} \frac{n!}{(n-y_i)!y_i!}\theta^{y_i}(1-\theta)^{n-y_i}p(\theta). \tag{A.11}$$

All that remains is to decide on a form of the prior distribution over θ. Since θ is the probability of a binary event, it will be useful to utilize a distribution that is defined over the unit interval. More importantly, it will be very useful if we choose a prior distribution which combines with a binomial likelihood in a useful way. A distribution that accomplishes both of these goals is the Beta distribution, Be(a, b),

$$\mathrm{Be}(a, b) = \frac{1}{\beta} x^{a-1}(1-x)^{b-1}. \tag{A.12}$$

Depending on how we choose the hyperparameters a and b, we can quantify any prior confidence we have about the value of θ. Alternatively, letting $a = b = 1$ results in a flat prior distribution over θ. Figure A.1 shows beta distributions of different values of a and b. Notice that this distribution provides a very flexible way for us to quantify any prior knowledge we might have, or we can adopt a flat prior.

The most useful outcome of using a Beta prior is that this distribution is *conjugate* to our Binomial likelihood. Returning to Bayes' rule, we now have a form for the both the likelihood and the prior in our model.

$$p(\theta|y_N) = \prod_{i=1}^{N} p(y_i|\theta)p(\theta) \tag{A.13}$$

$$p(\theta|y_N) = \prod_{i=1}^{N} \frac{n!}{(n-y_i)!y_i!}\theta^{y_i}(1-\theta)^{n-y_i}\frac{1}{\beta}\theta^{a-1}(1-\theta)^{b-1}. \tag{A.14}$$

We can remove some terms that don't depend on θ and we still retain a distribution that is proportional to the posterior distribution,

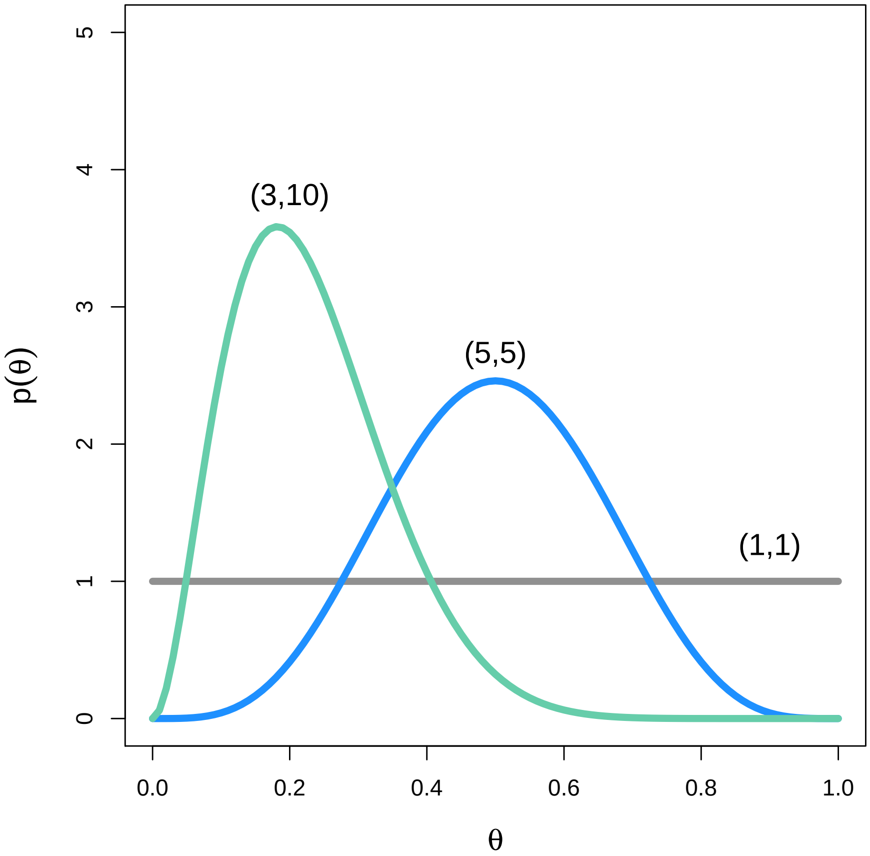

Figure A.1: The Beta distribution is shown with three different parameterizations.

$$p(\theta|y_N) \propto \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{n-y_i}\theta^{a-1}(1-\theta)^{b-1}. \tag{A.15}$$

It is now obvious that we can easily combine the components from the likelihood and the prior,

$$\begin{aligned} p(\theta|y_N) &\propto \prod_{i=1}^{N} \theta^{y_i+a-1}(1-\theta)^{n-y_i+b-1} && \text{(A.16)} \\ &= \theta^{\sum_i y_i+a-1}(1-\theta)^{\sum_i n-y_i+b-1}. && \text{(A.17)} \end{aligned}$$

Notice that this form of the posterior distribution has the same basic form as a Beta distribution. That is, the posterior distribution of θ is,

$$\begin{aligned} p(\theta|y_N) &\propto \text{Be}(A, B) && \text{(A.18)} \\ \text{where } A &= \sum_i y_i + a - 1 && \text{(A.19)} \\ B &= \sum_i n - y_i + b - 1. && \text{(A.20)} \end{aligned}$$

This is the primary benefit of thinking carefully about our prior distribution. If we pick a prior distribution which is conjugate to the likelihood, then the posterior will have the same form as the prior but with new parameters which are easily calculated from the data.

This example problem is continued in Figure A.2. In the left column are two simulated datasets drawn from Binomial distributions with $n = 4$ and

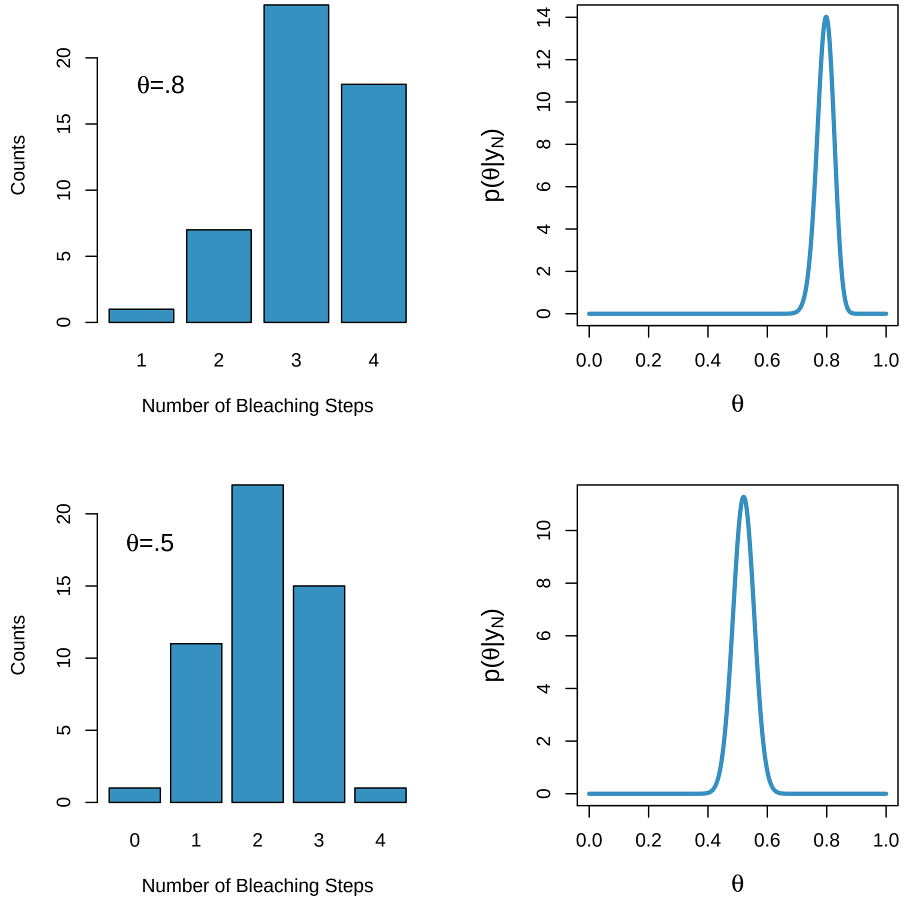

Figure A.2: Posterior estimation in the Beta-Binomial model. (Left) Samples drawn from a Binomial distribution with $n = 4$ and $\theta = .8$ (top) and $\theta = .5$ (bottom). (Right) The resulting posterior distributions of θ.

θ equal to 0.8 (top) and 0.5 (bottom). The right column shows the corresponding posterior distributions for θ. In this example, the hyperparameters of the prior distribution were both set to 1 which resulted in a flat prior distribution. Because of this flat prior, the peak of the posterior (MAP estimate) corresponds exactly to what we would estimate by maximizing the likelihood (ie. finding the best fit to the data). In addition to this point estimate, we also have a quantification of the whole parameter space and would easily be able to quantify parameter confidence and construct confidence intervals. Therefore, by choosing a conjugate prior, calculating the full posterior distribution over the parameters is achieved effortlessly.

I will describe one more example of a conjugate model which will also serve to transition us toward more generally applicable computational methods. Imagine that we have used patch clamp recording in order to measure the currents through a single ion channel. The transitions between *open* and *closed* states should follow Markovian dynamics, which prescribes that the duration of time spent in any state should be exponentially distributed (Colquhoun and Hakes, 1981). From our single channel recording, we tabulate the durations of each "dwell-time" and are left with a set of exponentially distributed random variables. It is our goal to estimate the corresponding timescale parameters of each distribution. Previous authors have thoroughly established successful methods for calculating these parameters using maximum likelihood methods (Colquhoun and Hawkes, 1981), but I describe the Bayesian way of approaching this problem.

Again, we imagine the data are drawn from an exponential distribution with unknown time scale parameter,

$$y_i \sim \theta e^{(-\theta y)}. \tag{A.21}$$

Given some data y_N, we want to estimate the posterior distribution over θ. Recalling Bayes' rule,

$$\begin{aligned} p(\theta|y_N) &\propto \prod_{i=1}^{N} p(y_i|\theta)p(\theta) && \text{(A.22)} \\ &= \prod_{i=1}^{N} \theta e^{(-\theta y_i)} p(\theta). && \text{(A.23)} \end{aligned}$$

Again, we want to carefully choose $p(\theta)$ so that it combines usefully with $p(y_i|\theta)$. The conjugate distribution to an Exponential likelihood is the Gamma distribution,

$$\mathrm{Ga}(a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{(-xb)}. \tag{A.24}$$

Combining likelihood and prior, we arrive at,

$$
\begin{aligned}
p(\theta|y_N) &\propto \prod_i p(y_i|\theta)p(\theta) && \text{(A.25)}\\
&= \prod_i \theta e^{(-\theta y)} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{(-\theta b)} && \text{(A.26)}\\
&\propto \prod_i e^{(-\theta y)} \theta^{a-1} e^{(-\theta b)} && \text{(A.27)}\\
&= \prod_i \theta^{a-1} e^{-\theta(y+b)} && \text{(A.28)}\\
&= \theta^{a+N-1} e^{-\theta(b+\sum_i y_i)}. && \text{(A.29)}
\end{aligned}
$$

We see that the posterior distribution of θ is a Gamma distribution with parameters that are easily calculated from the data.

$$
\begin{aligned}
p(\theta|y_N) &\propto \mathrm{Ga}(A, B) && \text{(A.30)}\\
\text{where } A &= a + N && \text{(A.31)}\\
B &= b + \sum y_i. && \text{(A.32)}
\end{aligned}
$$

I now extend this model into a more interesting case which will lead into our first computational method, Gibbs sampling. Instead of modeling the data as drawn from a single Exponential distribution, consider that we now imagine the data are drawn from a mixture of multiple Exponential distributions, a common case for single ion channel recordings. We imagine that each mixture has a distinct timescale parameter θ and mixture weight w. The data are drawn from some number of distinct components as,

$$y_i \sim w_1 e^{-\theta_1 y} + w_2 e^{-\theta_2 y} + ... + w_K e^{-\theta_K y} \tag{A.33}$$

$$= \sum_{j=1}^{K} w_j e^{\theta_j y}. \tag{A.34}$$

Without loss of generality, we focus on just a two-component exponential mixture for simplicity,

$$y_i \sim w_1 e^{-\theta_1 y} + w_2 e^{-\theta_2 y}. \tag{A.35}$$

The task then becomes estimating the four resulting free parameters from the data,

$$p(\theta_1, \theta_2, w_1, w_2 | y_N) \propto \prod_{i=1}^{N} p(y_i | \theta_1, \theta_2, w_1, w_2) p(\theta_1, \theta_2, w_1, w_2). \tag{A.36}$$

We wish to estimate a four-dimensional posterior distribution which spans the parameter space of the two timescale parameters and the two weight parameters. This kind of model will likely not have a simple closed form for the posterior, no matter how clever we may try to be with conjugate priors. However, we will be able to estimate the posterior distribution using a numerical method called Markov chain Monte Carlo (MCMC) sampling. The general strategy with MCMC is that while we may not be able to express a simple form for the posterior distribution, we could approximate its properties if we can draw a large number of independent and identically distributed (iid) samples from it. Importantly, even though we don't know the posterior distribution, we can draw iid samples by constructing a Markov chain whose limiting distribution is the posterior distribution. Then, by simply simulating this chain for many iterations, we draw many iid samples from the underlying distribution. Generating a Markov chain with a desired limiting distribution can be achieved in several ways, and I first describe Gibbs Sampling.

A.4 Gibbs Sampling

While we may not be able to devise a simple form for the posterior, $p(\theta_1, \theta_2, w_1, w_2|y_N)$, we can, with some care, devise a simple form for the *conditional* posterior of each parameter. As it turns out, this simple advance allows us to estimate the full posterior distribution using an MCMC algorithm called Gibbs Sampling (Geman and Geman, 1984; Gelfand and Smith, 1990). Before returning to this model, I will describe Gibbs sampling in general.

Consider a general joint probability distribution between two random variables, $p(A, B)$. From the definition of conditional probability,

$$p(A|B) = \frac{p(A, B)}{p(B)} \tag{A.37}$$

$$p(A, B) = p(B)p(A|B) \tag{A.38}$$

$$p(A, B) \propto p(A|B). \tag{A.39}$$

Similarly, we could calculate the condition density with respect to the other variable,

$$p(B|A) = \frac{p(A,B)}{p(A)} \tag{A.40}$$

$$p(A,B) = p(A)p(B|A) \tag{A.41}$$

$$p(A,B) \propto p(B|A). \tag{A.42}$$

Thus, the joint distribution, $p(A, B)$, is linearly proportional to both conditional distributions, $p(A|B)$ and $p(B|A)$. This fact holds generally for joint distributions over any number of random variables and is the basis of Gibbs Sampling. The strategy is that while the joint distribution, $p(A, B)$ might have no simple closed form, we can likely derive a simple form of each univariate conditional distribution. Generally, let $p(\theta_1, ..., \theta_K|x)$ be a K-dimensional posterior distribution with no simple closed form. If each univariate conditional distribution has a closed form such as $p(\theta_1|\theta_2, ..., \theta_K, x) \propto F(\theta_1)$, then Gibbs sampling proceeds by sequentially sampling each parameter conditioned on the previous samples of all other parameters. For each iteration i of the algorithm, we draw the i^{th} random sample of each parameter according the univariate conditional distributions,

$$\theta_1^i \sim p(\theta_1|\theta_2^{i-1}, ..., \theta_K^{i-1}, x) = F(\theta_1) \tag{A.43}$$

$$\theta_2^i \sim p(\theta_2|\theta_1^{i}, ..., \theta_K^{i-1}, x) = F(\theta_2) \tag{A.44}$$

$$... \tag{A.45}$$

$$\theta_K^i \sim p(\theta_K|\theta_1^{i}, ..., \theta_{K-1}^{i-1}, x) = F(\theta_K). \tag{A.46}$$

Therefore, being able to draw samples from each univariate conditional posterior allows us to construct a K-dimensional Markov chain which explores the parameter space in proportion to the posterior probability.

I now return to the two-component exponential mixture model. Recall that we decided we would be unable to devise a simple form for the four-dimensional posterior distribution, $p(\theta_1, \theta_2, w_1, w_2|y_N)$. However, we will see that it is straightforward to compute each conditional posterior, $p(\theta_1|\theta_2, w_1, w_2, y_N)$ and so on (for brevity, I now adopt the notation $p(\theta_1|...)$ to denote a conditional probability with respect to all other random variables in the model).

First, we employ a trick known as *data augmentation* by which we make the model more complicated in order to simplify the sampling scheme. In particular, I add new latent indicator variables $s_1, s_2, ..., s_N$ (one for each data point) which serve to label to which cluster a particular data point belongs. For our two component mixture model, each indicator variable points to one of the two mixture components, $s_i \in \{1, 2\}$. Our posterior distribution now has many parameters,

$$p(\theta_1, \theta_2, w_1, w_2, s_1, ..., s_N|y_N) \propto \prod_{i=1}^{N} p(y_i|...)p(\theta_1, \theta_2, w_1, w_2, s_1, ..., s_N), \quad \text{(A.47)}$$

but in the process of MCMC sampling, we *marginalize* out the latent variables s_i that we introduced,

$$p(\theta_1, \theta_2, w_1, w_2|y_N) = \int p(\theta_1, \theta_2, w_1, w_2, s_1, ..., s_N|y_N) ds_1 ds_2 ... ds_N. \qquad \text{(A.48)}$$

Therefore, even though we made the model more complicated by adding the s_i, we return to the desired model when we marginalize out the latent variables, which will be acheived with MCMC sampling of those parameters.

For simplicity, we can assume that the prior distribution for each parameter is independent,

$$p(\theta_1, \theta_2, w_1, w_2, s_1, ..., s_N|y_N) \propto \prod_{i=1}^{N} p(y_i|...)p(\theta_1)p(\theta_2)p(w_1)p(w_2)p(s_1)...p(s_N). \qquad \text{(A.49)}$$

To create our Gibbs Sampler, we need the conditional posterior distribution of each parameter, which is composed only of those components from the likelihood and prior that are relevant to each parameter. We seek,

$$p(\theta_j|...) = p(y_i|...)p(\theta_j) \qquad \text{(A.50)}$$

$$p(w_j|...) = p(y_i|...)p(w_j) \qquad \text{(A.51)}$$

$$p(s_i|...) = p(y_i|...)p(s_i). \qquad \text{(A.52)}$$

Relying on our previous results, we simply need to devise a conjugate prior for each parameter, and we will be able to easily sample from the corresponding conditional posterior. Now that we have the latent indicators s_i, let

A_j be the set of all i such that $s_i = j$. For each component, j, we already know a good conjugate model for estimating θ_j: the Exponential-Gamma model. Thus,

$$p(\theta_j|...) \propto \prod_{i \in A_j} w_j e^{-\theta_j y_i} \mathrm{Ga}(a, b) \tag{A.53}$$

$$= \mathrm{Ga}(a + |A_j|, b + \sum_{i \in A_j} y_i), \tag{A.54}$$

where $|A|$ denotes the number of elements in the set A. For each indicator variable, we need to sample s_i from the clusters $\{1, 2\}$ with probability equal to the posterior probability that data point i was drawn from each cluster. If we assume a flat prior on s_i, then this calculation boils down to calculating the likelihood that data point i was drawn from each cluster,

$$p(s_i = 1|...) \propto p(y_i|...)p(s_i) \tag{A.55}$$

$$\propto p(y_i|...) \tag{A.56}$$

$$= w_1 e^{\theta_1 y_i} \tag{A.57}$$

and

$$p(s_i = 2|...) \propto p(y_i|...)p(s_i) \tag{A.58}$$

$$\propto p(y_i|...) \tag{A.59}$$

$$= w_2 e^{\theta_2 y_i} \tag{A.60}$$

We then draw s_i from a Multinomial distribution with probability vector $\vec{p} = (p(s_i = 1), p(s_i = 2))$,

$$s_i \sim \text{Mult}(p(s_i = 1), p(s_i = 2)). \tag{A.61}$$

Thus, for each data point i we sample the indicator variable according to which component is likely to have generated y_i, conditioned on the current values of $\theta_1, \theta_2, w_1, w_2$.

The last part of our sampling scheme is the cluster weights w_j, for which we will encounter a new conjugate prior model. Note that the cluster indicator variables s_i are drawn from a Multinomial distribution and that the weights w_j for all the clusters must sum to 1. Consider the joint distribution of all the cluster weights (here, just two),

$$p(w_1, w_2|...) \propto p(y_i|...)p(w_1)p(w_2). \tag{A.62}$$

To sample the cluster weights, we take advantage of the conjugacy between a Multinomial likelihood and a Dirichlet prior. The Dirichlet distribution is a distribution over a vector of probabilities, which must sum to 1. A K-dimensional Dirichlet distribution is defined on the $(K-1)$-dimensional simplex, which ensures that the K elements drawn from this distribution will sum to 1. The Dirichlet distribution, with parameters $\alpha_1, ..., \alpha_K$ is,

$$\mathrm{Dir}(\alpha_1, ..., \alpha_K) = \frac{1}{\mathrm{B}(\vec{\alpha})} \prod_{j=1}^{K} x_j^{a_j - 1}. \tag{A.63}$$

This distribution, while perhaps unfamiliar, is easily seen as a generalization of the Binomial-Beta model we used earlier. In that instance, we were interested in a binomial likelihood which quantified the occurence of binary events. In particular, we wanted to know the parameter θ, the probability of a successful event. In that case, we had two possible outcomes (success or failure), each with probability θ and $(1-\theta)$, respectively. As the Multinomial distribution is a generalization of the Binomial to situations where we sample from many possible outcomes, the Dirichlet distribution is a generalization of the Beta, and quantifies the vector of probabilities of each outcome. Using this as a prior over the weights w_1, w_2 results in a Dirichlet posterior,

$$p(w_1, w_2|...) \propto \mathrm{Dir}(|A_1| + \alpha_1, |A_2| + \alpha_2). \tag{A.64}$$

With this, we have all the ingredients we need for our Gibbs Sampler. For each iteration of the algorithm, we draw random samples for each parameter as,

$$\theta_j|... \sim \mathrm{Ga}(a + |A_j|, b + \sum_{i \in A_j} y_i) \tag{A.65}$$

$$w_1, w_2|... \sim \mathrm{Dir}(|A_1| + \alpha_1, |A_2| + \alpha_2) \tag{A.66}$$

$$s_i|... \sim \mathrm{Mult}(p(s_i = 1), p(s_i = 2)). \tag{A.67}$$

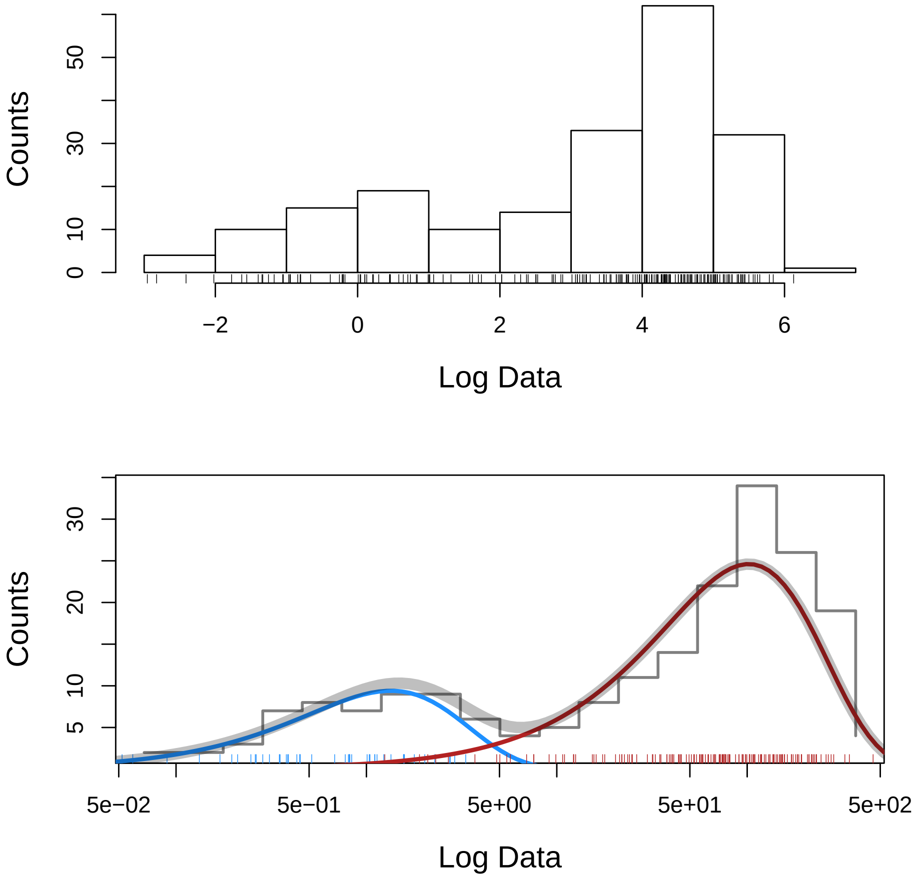

Figure A.3: A finite Exponential mixture model. (Top) Simulated data drawn from a mixture of two Exponential distributions. Data are plotted logarithmically for visualization. (Bottom) Result of using Gibbs sampler to infer the parameters of a two-component mixture model. Data points are colored corresponding to which component they are likely to have been generated from. The probability density of each component is shown and the sum of both densities is shown in gray and matches well with the histogram.

I next demonstrate this MCMC algorithm with simulated data. The data will be drawn from a mixture of two exponential distributions with timescale parameters $\theta_1 = 1$ and $\theta_2 = 100$ and with $w_1 = .25$ and $w_2 = .75$. The top of Figure A.3 is a histogram of the logarithm of each data point, and a rug plot of all the data is shown below. When visualized in this way, we can be sure that there are two distinct clusters within the data. One way to view the task of fitting these data is that we need to decide from which cluster each data point was drawn and then use the cluster assignments to estimate each θ_j and w_j. That is, we want the assignments of the s_i to yield high posterior probability. The Gibbs sampling scheme we just laid out will achieve this, and the result of this sampler is visualized in the bottom of Figure A.3. Here the data are shown again, but now each datapoint is labeled according to which cluster it is assigned: there is a blue cluster and a red cluster. These cluster labels correspond to just one iteration of the Gibbs sampler and thus represent a high posterior explanation of the data, but not the only one. Recall that we want to explore all the values of the parameters that yield good fits to the data, we want to explore the full posterior distribution. By sampling many cluster label assignments, all which yield high posterior probability, we marginalize out the s_i and yield accurate estimates of the *total* uncertainty in the model parameters that we're actually interested in.

Figure A.4 shows the result of our Gibbs sampler for each of the model parameters of interest: $\theta_1, \theta_2, w_1, w_2$. The top row shows the MCMC trajectories for the two dimensions of the Markov chain corresponding to the θ

parameters. Note that on the first iteration of MCMC, the parameters are initialized somewhere arbitrary in the parameter space, but quickly converge to a region of the parameter space that yields high posterior probability. This process of "burn-in" will be discussed in greater detail in the next section. After the Markov chain has converged, subsequent transitions yield iid samples from the posterior distribution. For each parameter, the positions of the chain can be aggregated together to approximate the marginal posterior distribution of each parameter, and this is shown in the second row of Figure A.4. This histogram of MCMC samples approximates the underlying marginal posterior and provides an accurate estimate of the parameter values and their uncertainty. Along with each histogram, the true parameter value is plotted as a vertical line and we see that our posterior distributions, from which we might construct a 95% confidence interval, accurately capture the underlying parameter values. The bottom half of Figure A.4 similarly shows the MCMC trajectories and marginal posterior distributions for the weight parameters w_1 and w_2. Again, we see that our MCMC estimate of the posterior distribution accurately captures the true parameter values and provides a natural way to quantify parameter confidence.

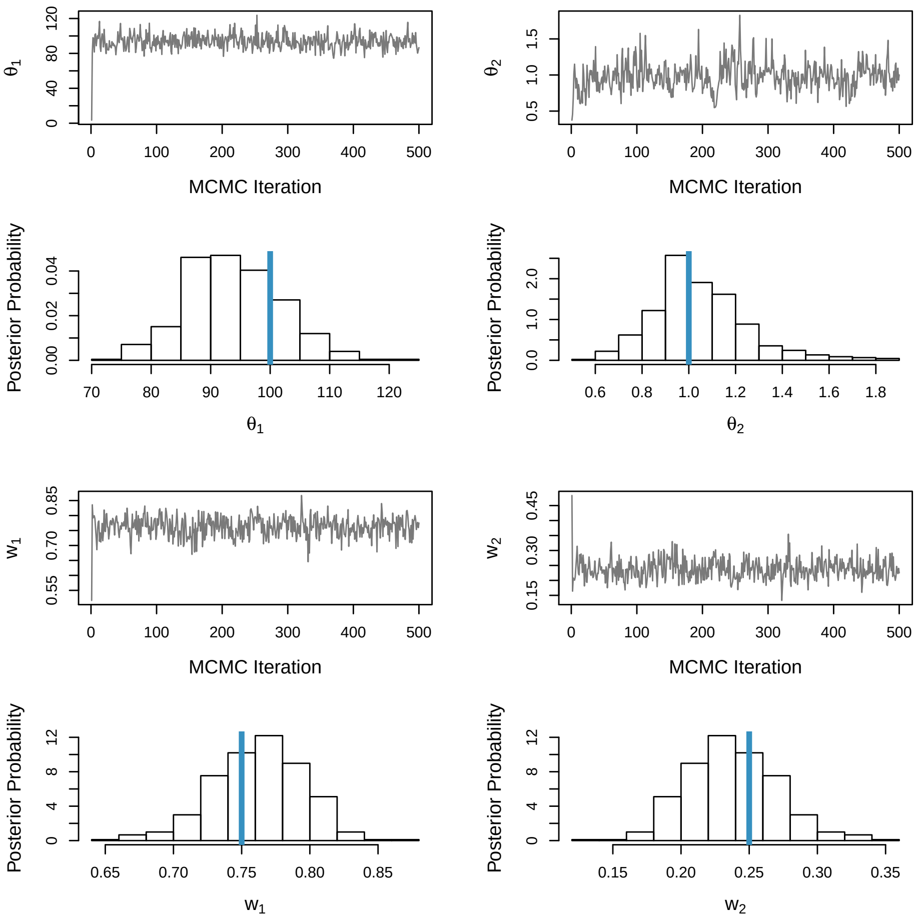

Figure A.4: Application of Gibbs sampling to the Exponential mixture data shown in Figure A.3. For each of the four model parameters $(\theta_1, \theta_2, w_1, w_2)$, the MCMC trajectories and marginal posterior distributions are shown.

Using MCMC, we are able to draw iid samples from a posterior distribution which is unknown to us, and therefore we can effectively estimate posteriors of any dimensionality. For Gibbs sampling, we only need a convenient form for each conditional posterior distribution, and the full posterior can be easily estimated. For many inference settings, this will be adequate as conjugate models have been devised for many kinds of distributions. Even very complicated probability models can be deconstructed into simple conditional posteriors for Gibbs sampling. For example, Hidden Markov Models tend to have many free parameters describing transitions dynamics and emission distributions (Rabiner, 1989). However, with useful data augmentation, the relevant conditional posteriors can be easily calculated and efficient Gibbs sampling scheme devised (Robert et al., 1993; Scott, 2002). This has already been applied in biophysical settings including modeling ion channel gating (Rosales, 2004; Siekmann et al., 2011). The Gibbs sampler, despite its simplicity and elegance, is inevitably limited to those models where we can calculate conditional posteriors. In some settings, this will not be possible and more general MCMC methods must be used.

A.5 Metropolis-Hastings

As a motivating example, I will consider the very general problem of curve-fitting. In common biophysical investigations, some theory is evaluated by its ability to explain data obtained from carefully controlled experimentation. Our model, with parameters $\theta_1, \theta_2, ..., \theta_K$ denoted $\vec{\theta}$, makes some prediction about how some measurable signal might look when examined with respect to some controlled variables. That is, our model prescribes some function $f(\vec{\theta}, x)$ which specifies how the observable signal f should depend upon model parameters $\vec{\theta}$ and independent variables x.

As a concrete example, imagine we are modeling the activation of a voltage-gated ion channel. A very simple model would be to assume the channel can exist in a conducting and non-conducting state, and the equilibrium between these states is perturbed by transmembrane voltage. Suppose we have measured a conductance-voltage (G-V) curve for this channel and want to fit it to a two-state Boltzmann distribution which quantifies the probability of the channel opening as a function of voltage. In this case, the independent variable is voltage, and our two-state model predicts that our G-V curve should follow the form,

$$f(a, b, V) = \frac{1}{1 + \exp(-V + a)^{b}}, \tag{A.68}$$

where parameters a and b might have some biophysical interpretation. Once we have made some measurements about how the channel activates at various controlled voltages, our goal is to find a good fit between the above equation and our data. That is, we want to fit the data by exploring the parameter space of a and b until the model prediction adequately matches the measured data. We might embark on this curve-fitting endeavor by searching the parameter space for the point which minimizes the error between the model and the data (Levenberg, 1944; Marquardt, 1963), and we would thus accept the resulting values of a and b as indicative of the true values of the underlying biophysical parameters. However, it has been noted that even for simple biophysical models, achieving a good fit to the data provides no guarantee that the recovered parameter estimates are accurate due to the pitfall of parameter non-identifiability (Hines et al., 2014). Therefore, we might prefer to take a Bayesian approach and seek not just a point estimate of parameters a and b, but instead to quantify the entire posterior distribution, $p(a, b|y)$.

In order to estimate the posterior distribution of our biophysical model, we will rely upon another MCMC methods called the Metropolis-Hastings algorithm. First, we decide that our observable signal, which is specified by our model in the form of some $f(\vec{\theta}, x)$ is also corrupted by the inevitable presence of experimental noise. For our example, we assume that $f(a, b, V)$ is accompanied by the presence of Normally distributed variability. This assumption isn't vital, any noise model could be used, but it seems reasonable in practice and is an assumption at the heart of existing curve fitting techniques such as error-minimization and maximum-likelihood (Seber, 2003b). That is, we assume that each data point y_i arises as a combination of a deterministic function $f(a, b, V)$ and some noisy process with unknown variance,

$$y_i \sim f(a, b, V_i) + N(0, \sigma^2). \tag{A.69}$$

Given this, our likelihood function is simply a Normal distribution centered at f and with variance σ^2,

$$p(y_i|...) = N(f(a,b,V_i),\sigma^2). \tag{A.70}$$

We assume that each data point arises from f and some iid noise, so the posterior distribution is,

$$p(a,b,\sigma^2|y_N) \propto \prod_{i=1}^{N} N(f(a,b,V_i),\sigma^2)p(a)p(b)p(\sigma^2). \tag{A.71}$$

When viewed in this way, we can start to guess that we won't be able to use Gibbs sampling here. It would not be straightforward to devise a conditional posterior, say $p(a|...)$, since our model parameters of interest are related to our likelihood only through a nonlinear function f. Therefore we have to turn to a more general method of MCMC.

Originally proposed by Nicholas Metropolis and colleagues to solve high dimensional problems in particle physics, what is now known as the Metropolis-Hastings algorithm is a very general tool for estimating probability distributions (Metropolis et al. 1953; Tierney, 1994). For simplicity, I'll describe only a special case of the Metropolis-Hastings method, called the Metropolis random walk. Recall that our posterior distribution of interest has three parameters: $\theta = \{a, b, \sigma^2\}$. We will construct a Markov chain whose limiting distribution is the posterior $p(a,b,\sigma^2|y_N)$. Using the Metropolis random walk, this Markov chain evolves with the following rules. At iteration i of the algorithm, the Markov chain is in location θ_i of the parameter space. We generate a *proposal* movement of the chain by taking a random walk from θ_i to a new location $\tilde{\theta}$. If the proposal point has higher posterior probability than θ_i (ie. if $p(\tilde{\theta}|...) > p(\theta_i|...)$), then we accept it and add it to the chain: $\theta_{i+1} = \tilde{\theta}$. If $p(\tilde{\theta}|...) < p(\theta_i|...)$, then we reject $\tilde{\theta}$ with probability α where α is the decrease in posterior probability: $p(\tilde{\theta}|y_n)/p(\theta_i|y_N)$. If the proposal is rejected, the Markov chain is extended with its current location, $\theta_{i+1} = \theta_i$. More succinctly, we can describe a single iteration of the simple Metropolis random walk algorithm as follows,

$$\begin{aligned}
&1. \quad \tilde{\theta} \sim \theta_i + N(\vec{0}, \Sigma) && \text{(A.72)}\\
&2a. \quad \text{if} \quad p(\tilde{\theta}|y_N) > p(\theta_i|y_N): \quad \theta_{i+1} = \tilde{\theta} && \text{(A.73)}\\
&2b. \quad \text{else draw} \quad u \sim U[0,1] && \text{(A.74)}\\
&\qquad \text{if} \quad u < \frac{p(\tilde{\theta}|y_N)}{p(\theta_i|y_N)}: \quad \theta_{i+1} = \tilde{\theta} && \text{(A.75)}\\
&\qquad \text{else}: \quad \theta_{i+1} = \theta_i && \text{(A.76)}
\end{aligned}$$

where Σ is a covariance matrix of our choice that specifies the characteristics of the random walk portion of the algorithm.

Let's break down, in a little more detail, what this algorithm does and how it works. The first component is that we attempt take a random walk in the parameter space, and if the proposal point leads to improved posterior probability then we keep it. This by itself would be a possible (albeit awfully slow) optimization method for finding the maximum of the posterior. But recall that the goal isn't to find a point estimate of the parameters, but instead to create a Markov chain that explores the whole parameter space in proportion to posterior probability. Thus, even if $\tilde{\theta}$ leads to a decrease in posterior probability, we still might keep it. And the probability with which we keep it is exactly the magnitude of the difference in posterior probability between θ_i and $\tilde{\theta}$. Suppose that θ_i is in an area of high posterior probability and that any random walk away from θ_i is likely to an area of lower posterior probability. We want the chain to be able to visit areas of lower posterior probability and this is eactly what the *accept/reject* rule achieves. If $p(\tilde{\theta}|y_N)$ is two-fold less than $p(\theta_i|y_N)$, then we only accept $\tilde{\theta}$ with probability $\frac{1}{2}$. And if $\tilde{\theta}$ is an area of much lower posterior probability, say 100-fold worse, then would only accept $\tilde{\theta}$ with probability $\frac{1}{100}$. In the algorithm above, we draw uniformly distributed random variables and compare them to $p(\tilde{\theta}|y_N)/p(\theta_i|y_N)$ as a particularly simple way of implementing this kind of accept/reject rule. Therefore the chain is able to explore all areas of the parameter space and not just areas of higher posterior probability than its current position. Further, the probability that the chain visits a particular location is exactly the posterior probability at that point in the parameter space, and we have successfully constructed a Markov chain whose limiting distribution is the posterior distribution.

It is important to appreciate what this algorithm has gained us. We decided that we would be unable to come up with a simple closed form for the desired posterior, $p(a, b, \sigma^2|y_N)$, or even any conditional distributions for

Gibbs sampling. Using the Metropolis random walk, we can estimate the posterior distribution for any model for which we can calculate the likelihood and the prior. This is a major advance. While we may not have a simple form for $p(y_i|\theta)p(\theta)$ for the whole parameter space, if we decide on a particular likelihood and prior, then it is trivial to compute the posterior probability for any particular parameter value θ_i, $p(y_i|\theta_i)p(\theta_i)$. In our example, we chose a Normal distribution for the likelihood and we can choose any kind of prior that we want for each parameter. Thus, we very easily walk around the parameter space, performing simple calculations of posterior probability and making accept/reject decisions and the result is iid samples from the posterior distribution.

Let's return to the example of G-V curves for a demonstration of the Metropolis random walk. At the top of Figure A.5 is a simulated activation curve generated with $a = -50$ and $b = 0.05$ and with added Gaussian noise with $\sigma = .02$. We can use this data to estimate the posterior distribution $p(a, b, \sigma^2|y_N)$ with the algorithm described above. In practice, we can implement the random walk in the full 3-dimensional space (as described above), or we can treat each parameter sequentially (within each iteration) and generate $\tilde{\theta}$ for a single parameter with a one dimensional random walk. Both approaches will work but there may be slight effects on chain mixing (see below) for some models. The result of MCMC is shown in Figure A.5 for the two parameters of interest, a and b. At left is the trajectory of each parameter of the course of MCMC and we see that while the parameters are initialized arbitrarily, they quickly converge to areas of higher posterior probability and explore only a small region of the parameter space. At right are histograms of the marginal posterior distributions along with the true parameter values plotted as vertical lines. Using the Metropolis random walk, we are able to easily recover an accurate estimate of the relevant parameters and their uncertainties. Importantly, to do this we only need to be able to calculate the expectation of the observable signal, $f(V|a,b)$, and the likelihood, $N(f|0,\sigma^2)$. Therefore, this approach can be used very generally in nearly all modeling endeavors.

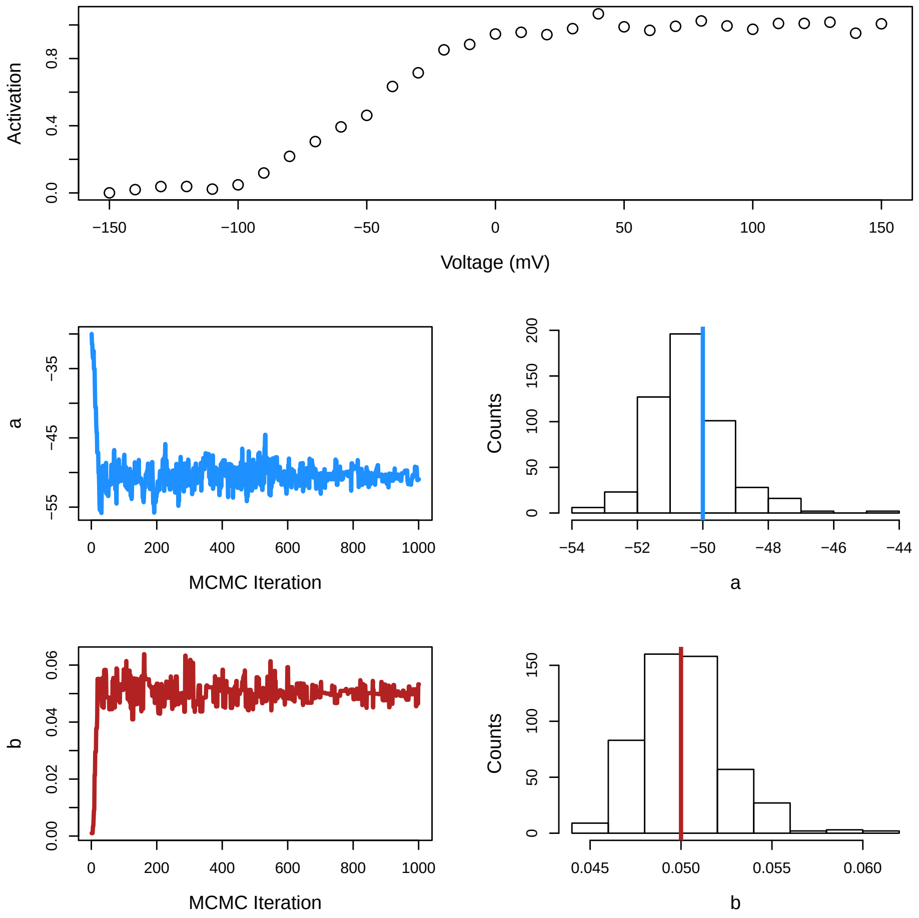

Figure A.5: Demonstration of Metropolis-Hastings algorithm to analyze ion channel activation data. (Top) Simulated G-V curve with added Gaussian noise. (Left) MCMC trajectories for model parameters a and b. (Right) Marginal posterior distributions of each parameter with the true values shown as vertical lines.

I now mention some practical details that need to be taken into consideration when implementing Metropolis-Hastings type methods. Let's revisit our example problem where we are trying to estimate parameters a and b. Notice that the parameter trajectories in Figure A.5 start off in bad areas of the parameter and move toward areas of high posterior probability where they eventually converge. The time period before this convergence is termed the burn-in as the Markov chain relaxes toward its actual limiting distribution. Parameter samples during the burn-in are discarded, and only samples from the true limiting distribution are iid samples from the posterior. This raises the important question of how to know when the chain has converged and begun providing legitimate iid samples from the posterior. Most simply we could just look at the trajectories - in Figure A.5, things certainly seem to have converged by 200 iterations, so that's probably a good cutoff. Naturally, more rigorous methods are desirable and many have been developed (Gelman and Rubin, 1992; Geweke, 1992), though I won't describe any in detail. Once we are confident (by whatever means) that the chain has converged, all subsequent motions of the chain yield iid samples from the posterior which we can use for parameter estimation.

There are important properties of the Metrpolis random walk that can affect chain mixing and convergence. For simplicity, let's suppose that our algorithm is implemented such that within each MCMC iteration, we generate a proposal $\tilde{\theta}$ and do accept/reject with each parameter sequentially. At each iteration, and for each parameter, we have the same basic algorithm, which for parameter a might look like,

$$\begin{aligned}
&1. \quad \tilde{a} \sim a_i + N(0,\eta) && \text{(A.77)}\\
&2a. \quad \text{if} \quad p(\tilde{a}|y_N) > p(a_i|y_N): \quad a_{i+1} = \tilde{a} && \text{(A.78)}\\
&2b. \quad \text{else draw} \quad u \sim U[0,1] && \text{(A.79)}\\
&\qquad \text{if} \quad u < \frac{p(\tilde{a}|y_N)}{p(a_i|y_N)}: \quad a_{i+1} = \tilde{a} && \text{(A.80)}\\
&\qquad \text{else}: \quad a_{i+1} = a_i. && \text{(A.81)}
\end{aligned}$$

The proposal points are drawn from a_i plus a Normally distributed random variable with standard deviation η. This Normal distribution, $N(0,\eta)$, which we might call the transition kernel of the random walk, can have an important impact on the Markov chain, which is explored in Figure A.6. Consider that η is very large, the case that is shown for the parameter trajectory at the top of Figure A.6. In this case, every proposal point $\tilde{a}$ is likely to be in a very different area of the parameter space than a_i. This has two major effects. First, notice that the Markov chain moves very quickly from the initial position to an area of high posterior probability in just a few iterations. This is because each iteration encompasses very large potential step sizes, since η is large. Having reached the posterior mode very quickly, the large step size is actually a detriment to posterior estimation. Consider that all subsequent proposals $\tilde{a}$ are likely to be very far away from a_i and will probably result in much lower posterior probability. It is therefore likely that $\tilde{a}$ will be rejected and the chain will stay in the same place. With large η, this rejection tends to happen iteration after iteration since many $\tilde{a}$ have very low posterior probability. The top trace in Figure A.6 shows long periods with no transitions and we get a small number of unique estimates of the parameter. The theory underlying MCMC guarantees that *any* Markov chain constructed with the Metropolis random walk will yield iid samples from the posterior *eventually* (Tierney, 1994), but with a poorly mixing chain we would have to simulate for a much longer time to get a useful sampling of the posterior.

We might be tempted to always avoid this problem by choosing η to be very small, but of course this strategy has its own drawbacks. This situation is shown in the middle trace of Figure A.6. We see that there are no long periods of rejections, in fact, there may be not a single rejection in this entire MCMC run. With small step sizes, all proposals $\tilde{a}$ are very close to a_i and thus have comparable posterior probability and are almost always accepted. But looking at this trace, we see the major shortcoming of η being too small. With small steps sizes, it takes an incredibly long time to move through the parameter space. In the middle trace, the chain barely makes it from the initial position to the posterior mode within the duration of the simulation. And once it reaches the area of high posterior probability, it moves very slowly through the parameter space and with high autocorrelation (something to be avoided). We are again in a situation where we would have to run an MCMC simulation for a very long time in order to get an adequate exploration of the posterior.

The solution, naturally enough, is some sort of indefinable in-between that depends on personal preference. A pretty good trace is shown at the bottom of Figure A.6. It moves quickly to the posterior mode, and once within the mode it still makes punctuated jumps around the parameter space separated by an adequate amount of rejection iterations. A through discussion of these practical considerations can be found in (Gilks et al., 1996a). Nonetheless, with adequate sampling, we can use the Metropolis-Hastings method very generally to aid in estimation problems for biophysical systems.

A.6 Conclusions

These Bayesian methods provide a very general paradigm for parameter inference in biophysics. With simple problems, we can calculate posterior distributions directly by using conjugate models. With more complex models, we can easily turn to computational methods for posterior inference, such as Gibbs sampling or the Metropolis-Hastings algorithm. Additionally, more sophisticated sampling methods exist which will be useful for exploring very high-dimensional posterior distributions (Neal, 2010; Girolami et al., 2011). The use of Bayesian methods for parameter inference gains us three advantages. First, it allows us to express parameter uncertainty as probability, a much more natural notion than that of the Frequentist sampling distribution. Second, we gain a simple mechanism to incorporate into the inference process any prior information we might have. Most importantly, Bayesian inference (with the aid of MCMC) gives us a generalizable method of rigorously addressing parameter inference and identifiability for arbitrarily complicated models.

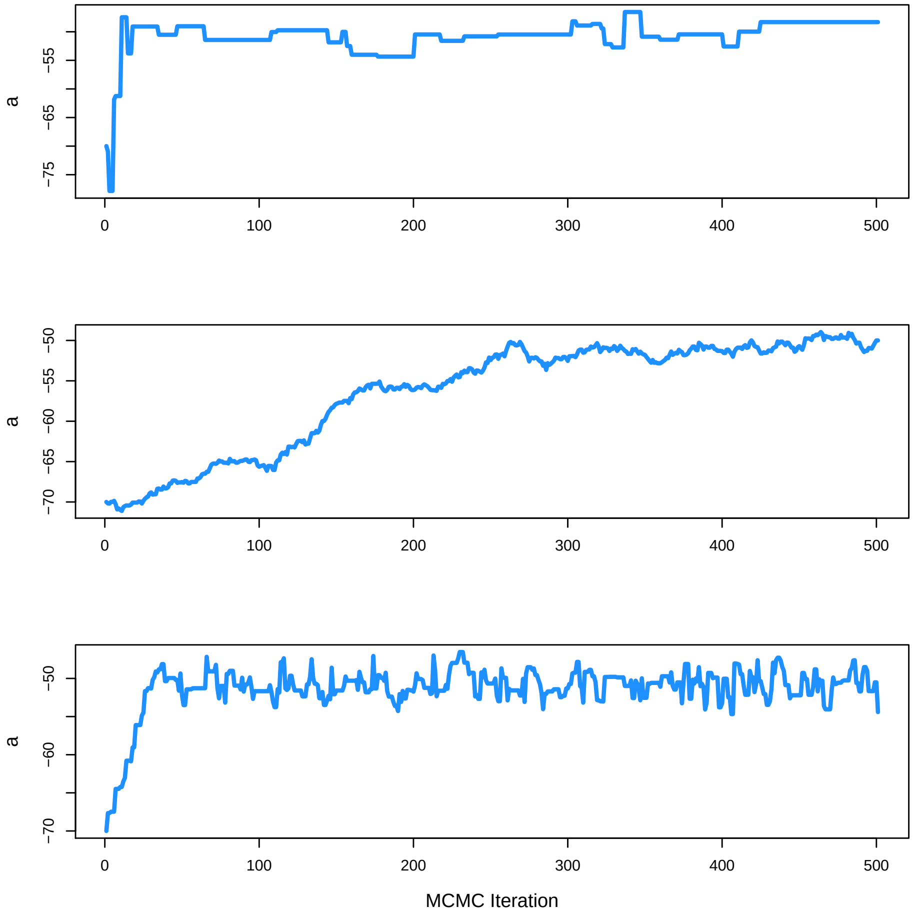

Figure A.6: These traces show the MCMC trajectory of a parameter with three different transition kernels. (Top) Transition step size is very large and though the chain moves quickly to the posterior mode, sampling is inefficient. (Middle) Transition step size is very small and chain moves very slowly through the parameter space. (Bottom) A compromise: chain moves quickly and with low autocorrelation.

Bibliography

Ackers, G., Doyle, M., Myers, D., and Daugherty, M. (1992). Molecular code for cooperativity in hemoglobin. *Science*, 225(5040):54–63.

Adair, G. (1925). The hemoglobin system iv: The oxygen dissociation curve of hemoglobin. *Journal of Biological Chemistry*, 63:529–545.

Armstrong, C. and Bezanilla, F. (1973). Currents related to movement of the gating particles of the sodium channels. *Nature*, 242:459–461.

Arumugam, S., Lee, T., and Benkovic, S. (2009). Investigation of stoichiometry of t4 bacteriophage helices loader protein (gp59). *Journal of Biological Chemistry*, 284:29283–29289.

Audoly, S., Bellu, G., D'Angio, L., Saccomani, M., and Cobelli, C. (2001). Global identifiability of nonlinear models of biological systems. *IEEE Transactions Biomedical Engineering*, 48:55–65.

Ball, F. and Sansom, M. (1989). Ion-channel gating mechanisms: model identification and parameter estimation from single channel recordings. *Proceedings of the Royal Society of London B*, 236:385–416.

Bankston, J., Camp, S., DiMaio, F., Lewis, A., Chetkovich, D., and Zagotta, W. (2012). Structure and stoichiometry of an accessory subunit trip8b interaction with hyperpolarization-activated cyclic nucleotide-gated channels. *Proceedings of the National Academy of Sciences*, 109(20):7899–904.

Baumgartner, W., Hohenthanner, K., Hofer, G., Groschner, K., and Romanin, C. (1997). Estimating the number of channels in patch-clamp recordings: application to kinetics analysis of multichannel data from voltage-operated channels. *Biophysical Journal*, 72:1143–1152.

Beal, M., Ghahramani, Z., and Rasmussen, C. (2002). The infinite hidden markov model. *Advances in Neural Information Processing Systems*.

Bellman, R. and Astrom, K. (1970). On structural identifiability. *Mathematical Biosciences*, 7:329–339.

Blei, D., Griffiths, T., Jordan, M., and Tennenbaum, J. (2004). Hierarchical topics models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems*.

Bottogtokh, D., Ash, D., Case, M., Arnold, J., and Schuttler, H. (2002). An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of neurospora crassa. *Proceedings of the National Academy of Sciences*, 99:16904–16909.

Brooks, S., Gelman, A., and Jones, G. (2011). *Handbook and Markov Chain Monte Carlo*. Chapman and Hall/ CRC Press.

Bruno, W., Wang, J., and Pearson, J. (2005). Using independent open-to-close transitions to simplify aggregated markov models of ion channel gating kinetics. *Proceedings of the National Academy of Sciences*, 102(18):6326–6331.

Burger, D., Cox, J., Comte, M., and Stein, E. (1984). Sequential conformational changes in calmodulin upon binding of calcium. *Biochemistry*, 23:1966–1971.

Calderhead, B., Epstein, M., Sivilotti, L., and Girolami, M. (2013). Bayesian approaches for mechanistic ion channel gating. In Schneider, M., editor, *In Silico Systems Biology*. Springer.

Caterall, W. (2012). Sodium channel mutations and epilepsy. In Noebels, J., Avoli, M., Rogawski, M., Olsen, R., and and, A. D.-E., editors, *Jasper's Basic Mechanisms of Epilepsies*. National Center for Biotechnology Information.

Celentano, J. and Hawkes, A. (2004). Use of covariance matrix in directly fitting kinetic parameters: application of gaba-a receptors. *Biophysical Journal*, 87:276–294.

Chappell, M. and Godfrey, K. (1992). Structural identifiability of the parameters of a nonlinear batch reactor model. *Mathematical Biosciences*, 108:241–251.

Cheung, S., Yates, J., and Aarons, L. (2013). The design and analysis of parallel experiments to produce structurally identifiable models. *Journal of Pharmacokinetics and Pharmacodynmics*, 40:93–100.

Cheung, W. (1980). Calmodulin plays a pivotal role in cellular regulation. *Science*, 207:19–27.

Cheung, W., Lynch, T., and Wallace, R. (1978). An endogenous ca2+-dependent activator protein of brain adenylate cyclase and cyclic nucleotide phosphodiesterase. *Advances in Cyclic Nucleotide Research*, 9:233–251.

Chis, O., Banga, J., and Balsa-Canto, E. (2011). Structural identifiability of systems biology models: A critical comparison of methods. *Plos ONE*, 6(1):1–16.

Cobelli, C. and Stefano, J. (1980). Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology*, 239:R7–R24.

Colquhoun, D. (1998). Binding, gating, affinity and efficacy: the interpretation of structure-activity relationships for agonists and of the effects of mutating receptors. *British Journal of Pharmacology*, 125:923–947.

Colquhoun, D., Hatton, C., and Hawkes, A. (2003). The quality of maximum likelihood estimates of ion channel rate constants. *Journal of Physiology*, 547:599–728.

Colquhoun, D. and Hawkes, A. (1981). On the stochastic properties of single ion channels. *Proceedings of the Royal Society of London B*, 211(1183):205–235.

Colquhoun, D. and Sakmann, B. (1985). Fast events in single-channel currents activated by acetylcholine and its analogues at the frog muscle end-plate. *Journal of Physiology*, 369:501–557.

Colquhoun, D. and Sigworth, F. (1983). Fitting and analysis of single-channel records. In Sakmann, B. and Neher, E., editors, *Single Channel Recording*. Plenum Press.

Coste, B., Xiao, B., Santos, J., Syeda, R., Grandl, J., Spencer, K., Kim, S., Schmidt, M., Mathur, J., Dubin, A., Montal, M., and Patapoutian, A. (2012). Piezo proteins are the pore-forming subunits of mechanically activated channels. *Nature*, 483:176–181.

Cox, D., Cui, J., and Aldrich, R. (1997). Allosteric gating of a large conductance ca-activated k+ channel. *Journal of General Physiology*, 110:257–281.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Crivici, A. and Ikura, M. (1995). Molecular and structural basis of target recognition by calmodulin. *Annual Reviews of Biophysics and Biomolecular Structure*, 24:85–116.

Crouch, T. and Klee, C. (1980). Positive cooperative binding of calcium to bovine brain calmodulin. *Biochemistry*, 19:3692–3698.

Csanady, L. (2006). Statistical evaluation of ion-channel gating models based on distributions of log-likelihood ratios. *Biophysical Journal*, 90:3523–3545.

Demuro, A., Penna, A., Safrina, O., Yeromin, A., Amcheslavsky, A., Cahalan, M., and Parker, I. (2011). Subunit stoichiometry of human orai1 and orai3 channels in closed and open states. *Proceedings of the National Academy of Sciences*, 108:17832–17837.

DiCera, E. (1995). *Thermodynamic Theory of Site-Specific Binding Processes in Biological Macromolecules*. Cambridge University Press.

Ding, H., Wong, P., Lee, E., Gafni, A., and Steel, D. (2009). Determination of the oligomer size of amyloidogenic protein β-amyloid(1-40) by single molecule spectroscopy. *Biophysical Journal*, 97:912–921.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Eydgahi, H., Chen, W., Muhlich, J., Vitkup, D., Tsitsiklis, J., and Sorger, P. (2013). Properties of cell death models calibrated and compared using bayesian approaches. *Molecular Systems Biology*, 9:664.

Eykhoff, P. (1974). *System Identification, Parameter and State Estimation*. Wiley.

Faller, D., Klingmuller, U., and Timmer, J. (2003). Simulation methods for optimal experimental design in systems biology. *Simulation*, 79:717–725.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–368.

Flomenbom, O. and Silbey, R. (2006). Utilizing the information content in two-state trajectories. *Proceedings of the National Academy of Sciences*, 103(29):10907 – 10910.

Forsen, S. and Linse, S. (1995). Cooperativity: Over the hill. *Trends in Biochemical Science*, 20:495–497.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2008). An hdp-hmm for systems with state persistence. *Proceedings of the 25th International Conference on Machine Learning.*

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2011). A sticky hdp-hmm with application to speaker diarization. *Annals of Applied Statistics*, 5(2A):1020–56.

Gelfand, A. and Smith, F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxations, gibbs distributions, and the bayesian restoration of image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Geweke, J. (1989). Bayesian inference in econometrics models using monte carlo integration. *Econometrica*, 57(6):1317–1339.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernardo, J., Berger, J., Dawiv, A., and Smith, A., editors, *Bayesian Statistics*. Clarendon Press.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996a). Introducing markov chain monte carlo. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov Chain Monte Carlo In Practice*. Chapman Hall Press.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996b). *Markov Chain Monte Carlo in Practice*. Chapman Hall Press.

Girolami, M. and Calderhead, B. (2011). Reimann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society B*, 73:123–214.

Godfrey, K., Jones, R., Brown, R., and Norton, J. (1982). Factors affecting the identifiability of compartmental models. *Automatica*, 18:285–293.

Goodwin, G. and Payne, R. (1977). *Dynamic System Identification: Experiment Design and Data Analysis*. Academic Press.

Grewal, M. and Glover, K. (1976). Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions Automatic Control*, 21:833–837.

Groulx, N., McGuire, H., Laprade, R., Schwartz, J., and Blunck, R. (2011). Single molecule fluorescence study of the *Bacillus thuringiensis* toxin cry1aa reveals tetramerization. *Journal of Biological Chemistry*, 286:42274–42282.

Gustavsson, I. (1975). Survey of applications of identification in chemical and physical processes. *Automatica*, 11:3–24.

Haiech, J., Klee, C., and Demaille, J. (1981). Effects of cations on affinity of calmodulin for calcium: ordered binding of calcium ions allows the specific activation of calmodulin-stimulated enzymes. *Biochemistry*, 20:3890–3897.

Hamill, O., Marty, A., Neher, E., Sakmann, B., and Sigworth, F. (1981). Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflugers Archive*, 391(2):85–100.

Hawkes, A., Jalali, A., and Colquhoun, D. (1992). Asymptotic distributions of apparent open times and shut times in a single channel record allowing for the omission of brief events. *Philosophical Transactions of the Royal Society of London B*, 337(1282):383–404.

Hengl, S., Kreutz, C., Timmer, J., and Maiwald, T. (2007). Data-based identifiability analysis of nonlinear dynamical systems. *Bioinformatics*, 23:2612–2618.

Higgs, D., Vickers, M., Wilkie, A., Pretorius, I., Jarman, A., and Weatherall, D. (1989). A review of the molecular genetics of the human alpha-globin gene cluster. *Blood*, 73(5):1081–104.

Hill, A. (1913). The combinations of haemoglobin with oxygen and with carbon monoxide. *Biochemical Journal*, 7:471–480.

Hines, K. (2013). Inferring subunit stoichiometry from singe molecule photobleaching. *Journal of General Physiology*, 141(6):737–746.

Hines, K., Middendorf, T., and Aldrich, R. (2014). Determination of parameter identifiability in nonlinear biophysical models: A bayesian approach. *Journal of General Physiology*, 143(3):401–416.

Hjort, N., Holmes, C., Mueller, P., and Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hodgdon, M. and Green, P. (1999). Bayesian choice among markov models of ion channels using markov chain monte carlo. *Proceedings of the Royal Society of London A*, 455:3425–3448.

Hodgkin, A. and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 103(2):321–62.

Hoefich, K. and Ikura, M. (2002). Calmodulin in action: Diversity in target recognition and activation mechanisms. *Cell*, 108:739–742.

Horn, R. (1987). Statistical methods for model discrimination: Applications to gating kinetics and permeation of the acetylcholine receptor channel. *Biophysical Journal*, 51(2):255–63.

Horn, R. and Lange, K. (1983). Estimating the kinetic constants from single channel data. *Biophysical Journal*, 43(2):207–223.

Horrigan, F. and Aldrich, R. (2002). Coupling between voltage sensor activation, ca2+ binding and channel opening in large conductance (bk) potassium channels. *Journal of General Physiology*, 120(3):267–305.

Jacquez, J. and Grief, P. (1985). Numerical parameter identifiability and estimability: integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77:201–227.

Jeffrys, H. (1945). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A*, 186(1007):453–61.

Jennrich, R. and Ralston, M. (1979). Fitting nonlinear models to data. *Annual Reviews of Biophysics and Bioengineering*, 8:195–238.

Ji, W., Xu, P., Li, Z., Lu, J., Liu, L., Zhan, Y., Chen, Y., Hille, B., Xu, T., and Chen, L. (2008). Functional stoichiometry of the unitary calcium-release-activated calcium channel. *Proceedings of the National Academy of Sciences*, 105:13668–13673.

Johnson, K., Simpson, Z., and Blom, T. (2009a). Fitspace explorer: An algorithm to evaluate multidimensional parameter space in fitting kinetic data. *Analytical Biochemistry*, 387:30–41.

Johnson, K., Simpson, Z., and Blom, T. (2009b). Global kinetic explorer: A new computer program for dynamic simulation and fitting of kinetic data. *Analytical Biochemistry*, 387:20–29.

Johnson, M. (2010). *Essential Numerical Methods*. Academic Press.

Johnson, M. and Faunt, L. (1992). Parameter estimation by least squares methods. *Methods in Enzymology*, 210:1–37.

Karlin, A. (1967). On the application of a plausible model of allosteric proteins to the receptor for the acetylcholine. *Journal of Theoretical Biology*, 16(2):306–320.

Keynes, R. and Rojas, E. (1974). Kinetics and steady-state properties of the charged system controlling sodium conductance in the squid giant axon. *Journal of Physiology*, 239:393–434.

Kienker, P. (1989). Equivalence of aggregated markov models of ion-channel gating. *Proceedings of the Royal Society of London B*, 236:269–309.

Kivinen, J., Sudderth, E., and Jordan, M. (2007). Learning multiscale representations of natural sciences using dirichlet processes. *IEEE Conference on Computer Vision*, pages 1–8.

Kline, D. (2009). An empirical bayesian approach for model-based inference of cellular signaling networks. *BMC Bioinformatics*, 10:371.

Klotz, I. (1997). *Ligand-Receptor Energetics: A Guide for the Perplexed*. Wiley.

Koopmans, T. (1949). Identification problems in economic model construction. *Econometrica*, 17:125–144.

Kusch, J., Thon, S., Schulz, E., Biskup, C., Nache, V., Zimmer, T., Seifert, R., Schwede, F., and Benndorf, K. (2012). How subunits cooperate in camp-induced activation of homotetrameric hcn2 channels. *Nature Chemical Biology*, 8:162–169.

Landowne, D., Yuan, B., and Magleby, K. L. (2013). Exponential sum-fitting of dwell-time distributions without specifying starting parameters. *Biophysical Journal*, 104:2383–2391.

Lape, R., Colquhoun, D., and Sivilotti, L. (2008). On the nature of partial agonism in the nicotinic receptor superfamily. *Nature*, 454(7205):722–7.

Lee, R. (1964). *Optimal Estimation, Identification and Control*. MIT Press.

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least-squares. *Quarterly Journal of Applied Mathematics*, 2:164–168.

Li, C. and Komatsuzaki, T. (2013). Aggregated markov models using time series of single molecule dwell times with minimum excessive information. *Physical Review Letters*, 111(058301):1–5.

Liebovitch, L. and Toth, T. (1990). The akaike information criterion (aic) is not a sufficient condition to determine the number of ion channel states from single channel recordings. *Synapse*, 5(2):385–416.

Linse, S., Helmersson, A., and Forsen, S. (1991). Calcium binding to calmodulin and its globular domains. *Journal of Biological Chemistry*, 266:8050–8054.

Ljung, L. (1987). *System Identification: Theory for the User.* Prentice Hall.

Lo, A. (1984). On a class of bayesian nonparametric estimates i: Density estimates. *Annals of Statistics*, 12:351–57.

Loredo, T. (1990). From laplace to supernova 1987a: Bayesian inference in astrphysics. In Fougere, P., editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers.

Marks, T. and Jones, S. (1992). An allosteric model for calcium channel activation and dihydropyridine action. *Journal of General*, 99:367–390.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics*, 11:431–441.

McGuire, H., Aurousseau, M., Bowie, D., and Blunck, R. (2012). Automating single subunit counting of membrane proteins in mammalian cells. *Journal of Biological Chemistry*, 287:35912–35921.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1954). Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Michaelis, L. and Menten, M. (1913). Die kinetik der invertinwirkung. *Biochemische Zietschrift*, 49:333–369.

Milescu, L., Akk, G., and Sachs, F. (2005). Maximum likelihood estimation of ion channel kinetics from macroscopic currents. *Biophysical Journal*, 88:2494–2515.

Millonas, M. and Hanck, D. (1998). Nonequilibrium response spectroscopy of voltage-sensitive ion channel gating. *Biophysical Journal*, 74(1):210–29.

Monod, J., Wyman, J., and Changeux, J. (1965). On the nature of allosteric transitions: a plausible model. *Journal of Molecular Biology*, 12:88–118.

Mueller, P. and Rodriguez, A. (2012). Nonparametric bayesian inference. In *Institute of Mathematical Statistics Monograph Series Volume 9*. Institute of Mathematical Statistics.

Muellner, F., Syed, S., Selvin, P., and Sigworth, F. (2010). Improved hidden markov models for molecular motors, part 1: Basic theory. *Biophysical Journal*, 99:3684–3695.

Nakajo, K., Ulbrich, M., Kubo, Y., and Isacoff, E. (2010). Stoichiometry of the kcnq1-kcne1 ion channel complex. *Proceedings of the National Academy of Sciences*, 107:18862–18867.

Neal, R. (2010). Mcmc using hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X., editors, *Handbook and Markov Chain Monte Carlo*. Chapman and Hall Press.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A*, 236:333–380.

Openheim, A. and Schafer, R. (1999). *Discrete-Time Signal Processing*. Prentice Hall.

Peersen, O., Madsen, T., and Falke, J. (1997). Intermolecular tuning of calmodulin by target peptides and proteins: Differential effect on ca2+ binding and implications for kinase activation. *Protein Science*, 6:794–807.

Pitman, J. (2002). Poisson-dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability, and Computing*, 11:501–514.

Porumb, T. (1994). Determination of calcium-binding constants by flow dialysis. *Analytical Biochemistry*, 220:227–237.

Qin, F., Auerbach, A., and Sachs, F. (1997). Maximum likelihood estimation of aggregated markov processes. *Proceedings of the Royal Society of London B*, 264(1380):375–383.

Rabiner, L. (1989). A tutorial on hidden markov models and select applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–86.

Ramaswamy, S., Cooper, D., Poddar, N., MacLean, D., Rambhadran, A., Taylor, J., Uhm, H., Landes, C., and Jayaraman, V. (2012). Role of conformational dynamics in alpha-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (ampa) receptor partial agonism. *Journal fo Biological Chemistry*, 287(52):43557–43564.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmueller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting profile likelihood. *Bioinformatics*, 25:1923–1929.

Reich, J. (1974). Analysis of kinetic and binding measurements i: Information content in kinetic data. *w*, 42:165–180.

Reich, J., Winkler, J., and Zinke, I. (1974a). Analysis of kinetic and binding measurements iv: Consistency of the mathematical model. *Studia Biophysica*, 42:77–99.

Reich, J., Winkler, J., and Zinke, I. (1974b). Analysis of kinetic and binding measurements iv: Reliability of parameter estimates. *Studia Biophysica*, 42:181–193.

Reich, J. and Zinke, I. (1974). Analysis of kinetic and binding measurements iv: Redundancy of model parameters. *J. Reich and I. Zinke*, 43:91–107.

Reiner, A., Arant, R., and Isacoff, E. (2012). Assembly stoichiometry of the gluk2/gluk5 kainate receptor complex. *Cell Reports*, 1:234–240.

Robert, C. and Casella, G. (2010). *Monte Carlo Statistical Methods.* Spring.

Robert, C., Celeux, G., and Diebolt, J. (1993). Bayesian estimation of hidden markov chains: A stochastic implementation. *Statistics and Probability Letters*, 16:77–83.

Roberts, R., Bobrow, M., and Bentley, D. (1992). Point mutations in the dystrophin gene. *Proceedings of the National Academy of Sciences*, 89:2331–2335.

Rosales, R. (2004). Mcmc for hidden markov models incorporating aggregation of states and filtering. *Bulletin for Mathematical Biology*, 66(5):1173–1199.

Rosales, R. and Varanda, W. (2009). Allosteric control of gating mechanisms revealed: the large conductance ca2+-activated k+ channel. *Biophysical Journal*, 96:3987–3996.

Rothberg, B. and Magleby, K. (2000). Voltage and ca2+ activation of single large-conductance ca2+-activated k+ channels described by two-tiered allosteric gating mechanism. *Journal of General Physiology*, 116:75–99.

Rotherberg, T. (1971). Identification in parametric models. *Econometrica*, 39:577–591.

Sakmann, B. and Neher, E. (1995). *Single Channel Recording.* Plenum Press.

Scott, S. (2002). Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351.

Seber, G. and Lee, A. (2003a). *Linear Regression Analysis*. Wiley Interscience.

Seber, G. and Lee, A. (2003b). *Nonlinear Regression*. Wiley Interscience.

Sethuraman, J. (1994). A constructive definition of dirichlet process priors. *Statistica Sinica*, 4:639–650.

Siekmann, I., Sneyd, J., and Crampin, E. (2012). Mcmc can detect nonidentifiable models. *Biophysical Journal*, 103(11):2275–86.

Siekmann, I., Wagner, L., Yule, D., Fox, C., Bryant, D., Crampin, E., and Sneyd, J. (2011). Mcmc estimation of markov models for ion channels. *Biophysical Journal*, 100(8):1919–29.

Sigworth, F. and Sine, S. (1987). Data transformations for improved display and fitting of single-channel dwell time histograms. *Biophysical Journal*, 52(6):1047–54.

Splawski, I., Shen, J., Timothy, K., Lehmann, M., Priori, S., Robinson, J., Moss, A., Schwarz, P., Towbin, J., and Vincent, G. (2000). Spectrum of mutations in long-qt syndrome genes. kvlqt1, herg, scn5a, kcne1, kcne2. *Circulation*, 102(10):1178–85.

Stigler, J. and Rief, M. (2012). Calcium-dependent folding of single calmodulin molecules. *Proceedings of the National Academy of Sciences*, 109:17814–17819.

Straume, M. and Johnson, M. (2010). Monte carlo method for determining complete confidence probability distributions of estimated model parameters. In Johnson, M., editor, *Essential Numerical Computer Methods*. Academic Press.

Svoboda, K., Schmidt, C., Schnapp, B., and Block, S. (1993). Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, 364:721–727.

Talukder, G. and Aldrich, R. (2000). Complex voltage-dependent behavior of single unliganded calcium-sensitive potassium channels. *Biophysical Journal*, 78(2):761–72.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1728.

Ulbrich, M. and Isacoff, E. (2007). Subunit counting in membrane-bound proteins. *Nature Methods*, 4:319–321.

Ulbrich, M. and Isacoff, E. (2008). Rules of engagement for nmda receptor subunits. *Proceedings of the National Academy of Sciences*, 105:14163–14168.

Vajda, S., Rabitz, H., Walter, E., and Lecourtier, Y. (1989). Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, 83:191–219.

van de Meent, J., Bronson, J., Wood, F., Gonzales, R., and Wiggins, C. (2013). Hierarchically-coupled hidden markov models for learning kinetic rates from single molecule data. *Journal of Machine Learning Research*, 28(2):361–369.

van Gael, J., Saatci, Y., Teh, Y., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden markov model. *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095.

Vankeerberghen, A., Wei, L., Jaspers, M., Cassiman, J., Nilius, B., and Cuppens, H. (1998). Characterization of 19 disease-associated missense mutations in the regulatory domain of the cystic fibrosis transmembrane conductance regulator. *Human Molecular Genetics*, 7(11):1761–1769.

Wagner, M. and Timmer, J. (2001). Model selection in non-nested hidden markov models for ion channel gating. *Journal of Theoretical Biology*, 208(4):439–50.

Walter, E. and Pronzato, L. (1997). *Identification of Parametric Models*. Springer.

Weiss, S. (2000). Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Structural Biology*, 7(9):724–729.

Wyman, J. and Gill, S. (1990). *Binding and Linkage: Functional Chemistry of Biological Macromolecules*. University Science Books.

Yap, K., Kim, J., Truong, K., Yaun, M. S. T., and Ikura, M. (2000). Calmodulin target database. *Journal of Structural and Functional Genomics*, 1:8–14.

Yu, Y., Ulbrich, M., Li, M., Dobbins, S., Zhang, W., Tong, L., Isacoff, E., and Yang, J. (2012). Molecular mechanism of the assembly of an acid-sensing receptor ion channel complex. *Nature Communications*, 3:1252.

Zagotta, W., Hoshi, T., and Aldrich, R. (1994). Shaker potassium channel gating iii: Evaluation of kinetic models for activation. *Journal of General Physiology*, 103(2):321–62.