Advancing Bayesian methods in fisheries stock assessment

Cole Monnahan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Trevor A. Branch, Chair

Ian J. Stewart

James T. Thorson

Program Authorized to Offer Degree:
Quantitative Ecology and Resource Management

University of Washington

Abstract

Advancing Bayesian methods in fisheries stock assessment

Cole Monnahan

Chair of the Supervisory Committee:
Associate Professor Trevor A. Branch
School of Aquatic and Fisheries Science

Inference is the process of drawing conclusions from data about unobserved quantities. Bayesian inference is one type of statistical inference and is widely applied in diverse fields. In fisheries, it has many advantages, notably a statistically rigorous way of including information from other studies (through prior distributions) and making probabilistic statements about key management quantities such as sustainable future levels of catch (in the estimated posterior distributions). Despite these advantages, it is rarely applied for integrated stock assessments due to computational hurdles (i.e., long run times). The goal of this dissertation is to advance computation methods for integrated models so that these methods can be more widely applied.

In chapter 1, I explore the potential of a new algorithm, the no-U-turn sampler (NUTS), to more efficiently sample the posterior distribution. Here, I compared the recently-developed Bayesian software Stan, which uses NUTS, to the most commonly used, JAGS, which belongs to the BUGS family of software. I found that NUTS was substantially more efficient, particularly as model size and complexity increased. However, hierarchical models were more sensitive to the parameterization with NUTS. I conclude that NUTS has high potential and should be incorporated in the software framework most commonly used in fisheries stock assessments, AD Model Builder (ADMB), and tested for stock assessments.

In my second chapter, I implemented NUTS into the source code of ADMB and com-

pared the efficiency of NUTS against the current Bayesian algorithm in ADMB, random walk Metropolis-Hastings (RWM) for six stock assessments, including an idealized, simulated model, four age-structured Stock Synthesis models, a custom-built, length-structured model, and a length- and age-based model used for research. I found that the main obstacle to fast run times was poorly parameterized models. One of the main causes of poor parameterization was overparameterized fishery selectivity curves, causing selectivity parameters to be near bounds, have long tails, and exhibit extreme correlation with other parameters. Most selectivity parameters are nuisance parameters that had no impact on management quantities, and so constraining the posterior with more informative priors and fixing parameters to the value at the bounds reduced run time by orders of magnitude while having negligible effect on model posterior distributions. An additional, and even more problematic, cause of poor parameterization was correlated early recruitment deviations, whose geometric shape challenged both NUTS and RWM algorithms. Most alarmingly, when models displayed these kinds of pathologies, the default RWM algorithm did not fully explore the posterior space even when apparently converged, which resulted in biased posterior samples. Even worse, this bias would not be detectable using traditional diagnostics, and longer RWM chains with more thinning would not help. In these cases, NUTS fared better, in that it was able to avoid the bias, but with greater accuracy came much slower run times. The end result of these explorations was a set of guidelines and the development of a software package designed to achieve run times 10-1000 times faster for most current stock assessment models.

In my last chapter, I examined the effect of hook spacing on Pacific halibut longline catch rates (CPUE) in commercial catch data. I found clear evidence for a hook spacing effect (i.e., hooks were less effective closer together) at the population level, using a spatially-explicit (geospatial) model with both non-parametric and parametric relationships. However, accounting for space had a greater impact on CPUE trends than did hook spacing, likely due to the relatively constant average hook spacing over time. Nevertheless, since constant

hook spacing is likely unusual in most fisheries over time, historical and future trends in hook spacing in commercial data can have important impacts on longline CPUE standardization. Accounting for hook spacing effects in other fisheries may improve the estimates of relative abundance trends, leading to better inference and thus management.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables viii

Introduction 1

Chapter 1: Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo 5

1.1 Abstract 5
1.2 Introduction 6
1.3 Principles of Hamiltonian Monte Carlo 8
1.4 Case studies 17
1.5 Results 20
1.6 Discussion 23
1.7 Acknowledgements 28

Chapter 2: Bayesian integration in fisheries stock assessments: confronting long run times 29

2.1 Abstract 29
2.2 Introduction 30
2.3 Methods 36
2.4 Results 45
2.5 Discussion 57
2.6 Acknowledgments 65

Chapter 3: The effect of hook spacing on longline catch rates: implications for catch rate standardization 66

3.1 Abstract 66

3.2 Introduction 67
3.3 Materials and Methods 72
3.4 Results 81
3.5 Discussion 86
3.6 Acknowledgements 89

Conclusion 94

Appendix A: Citation patterns of Bayesian software packages 112

Appendix B: Model files and code 113

Appendix C: Measuring MCMC efficiency 114

Appendix D: Efficient parameterizations for hierarchical models 116

Appendix E: Chapter 1 case study details 118
E.1 MVND and MVNC 118
E.2 Growth 119
E.3 Redkite 121
E.4 Swallows 121
E.5 Logistic 122
E.6 Wildflower 122

LIST OF FIGURES

Figure Number Page

1.1 Citation patterns of Stan and the BUGS family of Bayesian software platforms, for all journals in all fields. Data are from ISI Web of Science Core Collection. The y-axis units are the same, despite variable ranges. 7

1.2 Basics of Hamiltonian dynamics. (a) An example where a ball is dropped from the black point, it rolls down the surface over time (t), and momentum carries it up the other side where it reverses direction (red line), returning to where it started. The lines are offset to distinguish black and red paths. The position and momentum variables (b) and energies (c) over time corresponding to the path in (a). (d) Multiple paths for a 2d parabola. Gray dashed lines show posterior contours; initial positions and paths are red arrows and black lines. (e) Partial path (black line) on a posterior of a logistic population model with intrinsic growth rate (r) and carrying capacity (K). Red arrow shows initial position. (f) The energies for the trajectory in (e). 10

1.3 Examples demonstrating the basics of HMC. (a) The effect of different step sizes (ϵ) and number of steps (L) on trajectories. The blue and red trajectories approximate the same path (solid gray line), with the same initial position (red point) and trajectory length (ϵL), but opposite momentum. (b) Trajectories on a logistic posterior surface with identical initial position (black point) and momentum vectors. The black trajectory is slow to traverse the surface, while the red trajectory shows accumulating approximation errors, causing it to diverge. The blue trajectory utilizes a mass matrix, making the surface easier to traverse. (c) Multiple iterations of static HMC; black points are and accepted and intermediate steps (gray arrows) are discarded. (d) The acceptance ratios (α) of the trajectories in (b), with corresponding acceptance probability of $\min(1, \alpha)$. Multiple draws from the same initial position using a random walk Metropolis (e) or NUTS (f) algorithm, with and without an appropriate mass matrix (colors). 13

1.4 Comparison of efficiency (E) for Stan and JAGS across simulated models. The means (points) and ranges (segments) are across 20 replicates. (a) A multivariate normal with increasing dimensionality, either independent or with random correlations from an inverse-Wishart distribution. Ranges are too narrow to be visible. (b) A multivariate normal with repeated correlations on the off-diagonals for varying dimensions (MVNC). (c) A non-linear mixed effects model with two latent parameters per individual (Growth); ranges were left out for visual clarity. 21

1.5 Effects of noncentering on divergences and bias for the random effects on growth rate in the Growth model with 10 individuals. τ is the deviation from the mean for an arbitrary individual and the parameters in the centered model, σ its standard deviation, and $Z \sim N(0, 1)$ the parameters in the noncentered model. Samples from: (a) the centered model (target acceptance rate $\delta = 0.95$); (b) the noncentered model ($\delta = 0.80$); and (c) the transformed noncentered parameters, $\tau = \sigma Z$. Divergences in (a), shown in red, arise because the adapted step size is too large for the high gradients at low σ, creating an inaccessible region and leading to biased σ (i.e., no samples below $\log \sigma = -6$). The noncentered parameterization eliminates the curvature, and hence the divergences and bias (c). (d) Median rate of divergent transitions using $\delta = 0.80$ for both parameterizations. As information increases about σ (i.e., more individuals) the marginal distribution of σ narrows, simplifying the geometry and lowering the rate of divergences. 25

2.1 Citation patterns between ADMB and Stan across all fields through 2015. The y-axes have different ranges but the same scale. 38

2.2 The effect of the mass matrix on NUTS trajectories on a correlated bivariate normal model where each parameter is bounded between -1 and 1. Example NUTS trajectories (red lines) are generated from random points with random momenta, with 1000 independent posterior samples shown as black points. Rows indicate which type of mass matrix (M) is used, and columns the three different parameter spaces in ADMB. The algorithm works in the "unbounded" space, and inference is done in the "model space." The dense mass matrix is substantially easier to sample from as evidenced by the straighter trajectories in the bottom right panel. The x- and y-axis are the parameter values in each parameter space. 40

2.3 Diagnostics for the six slowest mixing **Cod** parameters, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses). 45

2.4 Diagnostics for the six slowest mixing **Hake** parameters, from an analysis of five pilot chains with 1000 samples each. Each column Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses). 46

2.5 Diagnostics for six poorly mixing **Canary** selectivity parameters, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses). 47

2.6 Diagnostics for six arbitrary **Snowcrab** parameters which have MLEs near bounds, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses). 48

2.7 Diagnostics for six consecutive early recruitment deviations in the **Halibut** model, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses). 49

2.8 Diagnostics for parameters from the **Tanner** model: the M parameter (first row/column), three non-consecutive recruitment deviations, and lastly the log of the posterior density (which has no MLE estimate). These are from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses). 50

2.9 Improvement in efficiency due to regularization for two case studies. Each point is associated with a parameter, and the lowest effective sample size is the bottleneck for each model. 51

2.10 Comparing efficiency between algorithms. Scatterplots of efficiency of RWM vs NUTS for each regularized model. Each point is a parameter, and those below the line suggest RWM mixes better, and vice versa for NUTS. 54

2.11 Samples from two adjacent recruitment deviations in the regularized Halibut model for the two algorithms. The NUTS algorithm is exploring a larger region of the posterior, suggesting that RWM is biased for this run. 55

2.12 Effect of sampling efficiency for both algorithms on regularized models for the default (MLE) mass matrix or an updated "dense" one. Parameters below the line mix better with the MLE mass matrix, while those above are improved by estimating a dense matrix. 56

2.13 Relative differences in parameter uncertainty between the two paradigms after regularizing. Frequentist estimates are standard errors and Bayesian estimates are marginal posterior variances (from RWM). Values below the line indicate that frequentist uncertainties are smaller. 57

2.14 Comparison of estimates of key management quantities for 3 regularized models. Posterior distributions are shown as gray histogram, with posterior median as red vertical line. The asymptotic estimate from the delta method, assumed to be normal, is shown as a black curve with the estimate a vertical black line. Different management qualities are relevant among models due to differences in setup. Spawning stock biomass (SSB), depletion (biomass relative to unfished state), and maximum sustainable yield (MSY) are common management metrics on the U.S. West Coast. 58

2.15 The double-normal selectivity pattern for the first fleet in the Canary model for the original with pathologies (see Fig. 2.5) and regularized versions of the model. The implied prior (top row) and posterior (bottom row) are shown as shaded regions, while the red line denotes the MLE. 61

3.1 Stylized representations of three hypotheses for how the power of a hook changes with hook spacing, for a set with the same number of hooks but increasing total length and thus hook spacing. In the *hook* hypothesis, the hooks do not compete and thus the effective effort is the nominal hooks. In the *length* hypothesis, hooks compete at all spacings such that the length of the set is the effective effort. Lastly, the *spacing* hypothesis is intermediate and hooks compete at lower spacings only. Figure recreated from [126]. . . . 69

3.2 Properties of the fishery-dependent data (commercial catches). (a) The distribution of hook spacing within each of the three gear types. (b) Trends in proportion of catches by gear type by weight. (c) Annual distribution of hook spacing for all gear types (small points; jittered for clarity) and means (large points). 71

3.3 The raw data from Hamley and Skud (1978). Each panel is a separate site, and each line represents a series of sets fished at different spacings on the same day. Day number is colored. Sets with zero catch are removed. 79

3.4 Estimated hook spacing effects for the smoother (a) and parametric forms from the spatiotemporal (b) and experimental model (c). Lines and shaded region show estimates and approximate 95% confidence interval, and red line shows historical parametric fit to experimental data from (Hamley and Skud 1978). 82

3.5 Effect of spatial component (rows) and hook spacing form (columns) on trends in relative abundance. Each panel is normalized by dividing by its mean. Lines and shaded region show estimates and approximate 95% confidence interval. 84

3.6 The relative abundance trend from the fishery-independent survey (blue) conducted under a controlled design (constant gear on a uniform grid, see [145]), compared to the parametric fit in the spatiotemporal model. Both are normalized to have mean of one. Lines and shaded region show estimates and approximate 95% confidence interval. 1996 and 1997 were unavailable for the survey series and thus left off for the spatiotemporal results. 86

LIST OF TABLES

Table Number Page

1.1 Summary of case studies used to compare efficiency between Stan and JAGS. Further details are available in Appendix E. Latent parameters are those modeled as random effects. 19

1.2 Case study results comparing efficiency of Stan and JAGS. Max correlation is the largest absolute pairwise correlation, calculated from converged samples. Efficiency (E) is the number of effective samples per time. 22

1.3 Summary of key differences between JAGS and Stan. 27

2.1 Summary of case studies used. Speed (ms/evaluation) is calculated from random walk Metropolis (RWM) runs in which gradients are not calculated. . . 42

2.2 Regularization and resulting changes to maximum likelihood estimates and standard errors (in parentheses) of select management quantities for the two successfully regularized models. Spawning stock biomass (SSB), depletion (biomass relative to unfished state), overfishing limit (OFL) and maximum sustainable yield (MSY) are common management metrics on the U.S. West Coast. 52

2.3 Effective samples for a 12hr (overnight) run using 10 parallel chains using RWM, and the ratio of efficiencies (E) between NUTS and RWM, where applicable. 53

3.1 Model estimates and standard errors (parentheses) for the parametric model fit to the experimental data. 82

3.2 Key model estimates and standard errors (parentheses) for models with and without space, and the parametric and smoother form for hook spacing. Depending on the model structure some parameters are not estimated, represented by (-), or the first level of a factor set to zero and thus there is no standard error. See appendix A for further results. 85

ACKNOWLEDGMENTS

For me, graduate school was about growing from a student to scientist and I have many people to thank. First, my advisor Trevor Branch was instrumental to this process. He gave me the time, freedom, and opportunities to develop my own research topics and approach to addressing them. I have enjoyed my time collaborating with him, and look forward to future projects as colleagues.

I am also grateful for the guidance and mentorship provided by my committee. I worked closely with Ian Stewart, and must thank both him and his employer, the International Pacific Halibut Commission, which shared data and let me partake in the assessment, but also sent me on the setline survey and to the annual meeting. To see the management process from data collection and processing, conducting the assessment, and finally dissemination of the results to the stakeholders was truly a unique and valuable opportunity. Without the guidance and interactions with Jim Thorson, I would not have learned half of the technical skills I did. I thank him for introducing and teaching advanced topics, but also individual help and ideas on the more technical aspects of my research. I also thank Tim Essington for helpful feedback on drafts and big picture context, and Miles Logsdon for being a responsive and supportive GSR. Lastly, I thank the numerous scientists that helped me in many aspects of my research. Namely, fishery scientists Dave Fouriner, Kasper Kristensen, Hans Skaug, Allan Hicks, and Ian Taylor. I would also like to thank members of the Stan development team: Bob Carpenter, Michael Betancourt, and Jonah Gabry were responsive and helpful in understanding HMC and Stan.

QERM has been the perfect fit for my skillset and interests, and I appreciate the unique nature of the program. I am proud to be a QERM alumni. Although a QERM student, I

spent my time at the UW housed in the Branch lab in SAFS. My lab mates, and the broader SAFS community, are an incredible group: brilliant scientists, but also diverse, collaborative, and socially engaging. I have met many lifelong collaborators, colleagues, and friends, and for that I am extremely grateful.

Lastly, I would like to thank the funding sources that made this research possible: the UW Joint Institute for the Study of the Atmosphere and Ocean and Washington Sea Grant Population Dynamics Fellowship.

DEDICATION

I dedicate my dissertation to my family, whose unwavering support made this long, arduous journey possible. I want to thank my father, Patrick, for instilling in me a productive work ethic, and my mom, Lynne, for her endless curiosity and support in my studies. Two of my biggest supporters for graduate school, grandmother Carole and uncle Thom were not here to see me finish, but I know how proud they would be. Finally, to my best friend and lovely wife, Deonna, who I thank for her encouragement and patience, particularly during the hard times. I am excited to start the next chapter of our lives together.

INTRODUCTION

The field of quantitative fisheries stock assessment has a long and rich history of theoretical development and practical application. Stock assessment models began as simple models considering only a single indicator of population size (often called biomass dynamics models), which aggregate processes like growth, recruitment, and gear selectivity across age, size, and sex [1]. The classic biomass dynamics model is a logistic growth model modified to include catch which is proportional to harvest ratio and stock size [2]. As the field advanced, more biological complexity was incorporated into the models, such as age structure, stock-recruit relationships, growth, as well as fishery properties like selectivity [1]. A major advancement was the development of the theory for statistical catch-at-age models [3, 4], including the approach of incorporating multiple data sources into a single analysis – known as integrated analysis [4, 5].

Many of these advances were possible due to the increasing hardware capabilities of computers, the development of powerful software libraries such as automatic differentiation [6] and flexible modeling frameworks such as AD Model Builder [7]. With these new hardware and software advances, analysts can write their own custom assessment models, with complexity tailored to the available data and nuances of each fishery. However, some analysts developed generalized, flexible stock assessment platforms that could be quickly adapted and applied to fisheries with a variety of data types and characteristics. Today, these generalized models are considered state of the art for data rich fisheries stock assessment modeling and represent a culmination of decades of work in theoretical models, software development, and hardware advances. The most commonly used model on the US West Coast is Stock Synthesis [8], although others exist with similar functionality, such as Coleraine [9], MULTIFAN-CL

[10], and CASAL [11].

In parallel with software developments, advancements have also been made in methods for fitting the models to data and making inference. The statistical theory for maximum likelihood and Bayesian inference has existed for decades, but the computational requirements to apply these methods to fisheries were initially an impediment. For example, the original catch-at-age models of [3] were fit using least squares applied to linear approximations of the catch equations, while today it is common to estimate hundreds (if not thousands) of parameters with standard errors (Ian Taylor, NWFSC, pers. comm.) using maximum likelihood.

Bayesian inference is an alternative framework to the maximum likelihood (frequentist) approach for conducting statistical inference, although they both share the same likelihood function. The fundamental difference is in the interpretation of a probability: a frequentist probability is the proportion of times an event occurs in an infinite sequence of repeated trials; a Bayesian probability is interpreted as a reasonable expectation of a degree of belief. Practically, frequentist fixed effect parameters are non-random and fixed but unknown, while Bayesian parameters (and derived quantities) are random variables with probability distributions [12]. Although Bayesian inference was introduced in 1774 [13], it was not until computational advances in the 1990s that solutions to applied fisheries problems became feasible (e.g., [14]).

This sparked a new wave of research in assessment models such as state-space surplus production [15, 16] and age-structured assessments [17], in addition to meta-analyses to determine informative biological priors [18]. These early papers made convincing arguments of the advantages of Bayesian methods and introduced the necessary algorithms with tractable examples, while also noting the need for further research on biological priors and numerical methods. Since then research and development of Bayesian analysis in many fields has led to an explosion of new techniques, algorithms, and software, and the growth in popularity of Bayesian methods (e.g., [19]). Besides computational methods, studies for informative biological priors have been conducted for some parameters (e.g., [18, 20]), but the availability

and implications of priors has not been thoroughly explored.

Another important advancement was in the development and implementation of hierarchical models (i.e., models with random effects). These models allow an explicit representation of data into observation and state processes [21] and provide a generic solution to non-independence caused by latent states [22]. These models are used extensively in a wide variety of fields, and early implementations in fisheries were for conducting meta-analyses (e.g., [18, 23]), estimating growth external to the assessment [24, 25] and state-space implementations of simple biomass dynamics models [15, 16].

In modern integrated models like Stock Synthesis, non-stationary processes such as recruitment, selectivity, and growth, are typically approximated by estimating annual deviations around an average in a penalized likelihood framework [8, 26]. Treating these deviations as random effects allows the data to provide information on the magnitude of stochastic processes. However, the computational challenge of integrating across the random effects in large, non-linear mixed effects models remains an impediment [5]. The Bayesian approach provides a natural solution since it automatically integrates across all parameters, while for non-Bayesian models, a recently developed approach using the Laplace approximation may also perform well in some cases [26]. Bayesian hierarchical modeling is thus a vital tool in fisheries stock assessment, whether providing informative priors, processing input data, or formally incorporating and accounting for non-stationary processes in the assessment models themselves.

Regardless of the model used, the goal of quantitative fisheries stock assessment is to evaluate the consequences of alternate management actions and ensure that catches do not result in biomass declining below target thresholds [1]. The true state of a stock is never known and management action may depend on alternative hypotheses, so there is a need to quantify the evidence of these hypotheses, formally known as decision analysis. The Bayesian approach offers a simple and elegant framework for estimating probabilities of hypotheses and is a natural way to perform decision analyses [27, 28].

While the Bayesian approach offers key advantages to stock assessment, there are signif-

icant technological issues preventing these analyses for many stocks. The primary hurdle is prohibitively long run times to achieve convergence, often several orders of magnitude too long for practical implementation. The goal of this dissertation is to advance the current state of Bayesian methods for complex, integrated fisheries stock assessments so that these methods can be used more broadly. In chapter 1 I investigate the potential for a new algorithm, the no-U-turn sampler [29], to make run times faster for a variety of hierarchical Bayesian models. I provide a basic illustration of the principles on a simple set of models to help analysts develop intuition about the complex nature of this algorithm, highlighting key concepts. In chapter 2 I turn my attention to stock assessment models with the goal of reducing run time by several orders of magnitude. Specifically, I explore what causes long run times in models, and implement and apply the no-U-turn sampler in ADMB. My third chapter uses hierarchical (Gaussian process) models to explore the effect of hook spacing on CPUE standardization for Pacific halibut. The resulting time series is useful as a data input to the stock assessment for this species.

Chapter 1

FASTER ESTIMATION OF BAYESIAN MODELS IN ECOLOGY USING HAMILTONIAN MONTE CARLO

1.1 Abstract

Bayesian inference is a powerful tool to better understand ecological processes across varied subfields in ecology, and is often implemented in generic and flexible software packages such as the widely-used BUGS family (BUGS, WinBUGS, OpenBUGS, and JAGS). However, some models have prohibitively long run times when implemented in BUGS. A relatively new software platform called Stan uses Hamiltonian Monte Carlo (HMC), a family of Markov chain Monte Carlo (MCMC) algorithms which promise improved efficiency and faster inference relative to those used by BUGS. Stan is gaining traction in many fields as an alternative to BUGS, but adoption has been slow in ecology, likely due in part to the complex nature of HMC.

Here, I provided an intuitive illustration of the principles of HMC on a set of simple models. I then compared the relative efficiency of BUGS and Stan using population ecology models that vary in size and complexity. For hierarchical models, I also investigated the effect of an alternative parameterization of random effects, known as noncentering.

For small, simple models there is little practical difference between the two platforms, but Stan outperforms BUGS as model size and complexity grows. Stan also performs well for hierarchical models, but is more sensitive to model parameterization than BUGS. Stan may also be more robust to biased inference caused by pathologies, because it produces diagnostic warnings where BUGS provides none. Disadvantages of Stan include an inability to use discrete parameters, more complex diagnostics, and a greater requirement for hands-on tuning.

Given these results, Stan is a valuable tool for many ecologists utilizing Bayesian inference, particularly for problems where BUGS is prohibitively slow. As such, Stan can extend the boundaries of feasible models for applied problems, leading to better understanding of ecological processes. Fields that would likely benefit include estimation of individual and population growth rates, meta-analyses and cross-system comparisons, and spatiotemporal models.

1.2 Introduction

Bayesian inference is used widely throughout ecology, including population dynamics, genetics, community ecology, and environmental impact assessment, among other subfields [19]. In the Bayesian paradigm, the likelihood of the observed data is combined with prior distributions on parameters, resulting in a posterior probability distribution of parameters, from which inference is made [12]. Expectations of posterior quantities, such as means or quantiles, are commonly approximated using numerical techniques, with Markov chain Monte Carlo (MCMC) being the most common [30]

The popularity of Bayesian inference grew particularly fast with the development of generic and flexible software platforms, with the BUGS family (here defined as BUGS, WinBUGS, OpenBUGS and JAGS; see Appendix A) being by far the most common (Fig. 1.1). For a given model, BUGS automatically selects an MCMC algorithm, and arguments controlling its behavior (i.e., tuning parameters), where necessary. The analyst can thus focus on the model and scientific questions, rather than the mechanics of the underlying MCMC algorithms. As such, these platforms have been the workhorse for Bayesian analyses in ecology and other fields for the last 20 years.

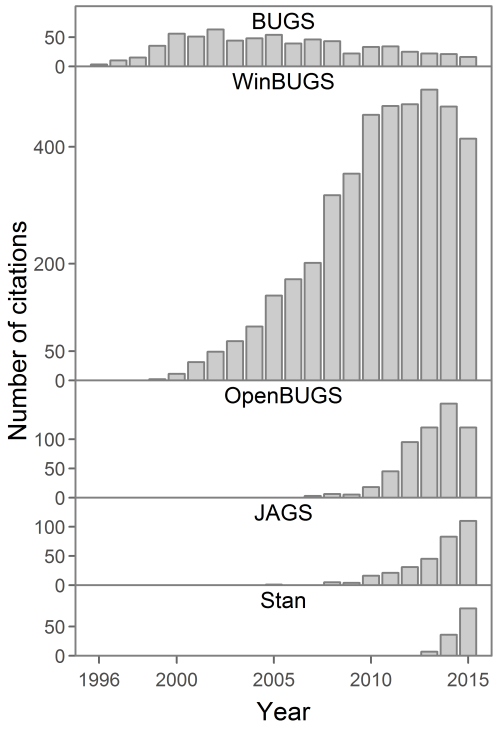

Figure 1.1: Citation patterns of Stan and the BUGS family of Bayesian software platforms, for all journals in all fields. Data are from ISI Web of Science Core Collection. The y-axis units are the same, despite variable ranges.

However, for certain models the time required for inference (runtime) using BUGS is prohibitively long. Long runtimes often occur in BUGS because the underlying MCMC algorithms are inefficient, which is further compounded when the model needs to run many times during development, model selection (e.g., cross validation; [31]), or simulation testing. These issues remain despite the increasing power of computers because data sets are increasing in size and models are becoming more complex [32]. At the same time, hierarchical modeling is becoming increasingly popular, as this type of model is widely recognized as a natural tool for formulating and thinking about problems in many ecological subfields [21, 33, 22]. Thus, there is a need for alternatives to BUGS that are faster across a range of model size, complexity, and hierarchical structure.

A family of MCMC algorithms called Hamiltonian Monte Carlo (HMC; [34]) promises improved efficiency over the algorithms used by BUGS, but until recently have been slow to be adopted for two reasons. First, HMC requires precise gradients (i.e., derivatives of the log-posterior with respect to parameters), but analytical formulas are rare and numerical techniques are imprecise, particularly in higher dimensions. Second, the original HMC algorithm requires expert, hands-on tuning to be efficient [34]. Both of these hurdles have recently been overcome, the first with automatic differentiation (e.g., [6]), and the second with an HMC algorithm known as the no-U-turn sampler (NUTS; [29]). These advances have been packaged into the open-source, generic and flexible modeling software Stan [35, 36, 37], which effectively aims to replace the BUGS family and is quickly gaining traction across diverse fields (Fig. 1.1).

Despite the potential of HMC, and the availability of Stan, adoption has been slow in ecology, likely because ecologists are either unaware of its existence, or are unsure when it should be preferred over BUGS. Here, I illustrate the principles that underlie HMC, and then compare the efficiency between Stan and a BUGS variant, JAGS [38], across a range of models in population ecology. Specifically, I test how HMC performance scales with model size and complexity, and its suitability for hierarchical models. My goal is explore the relative benefits of Stan and JAGS, and provide guidance for ecologists looking to use the power of HMC for faster and more robust Bayesian inference.

1.3 Principles of Hamiltonian Monte Carlo

The existing literature on HMC tends to focus on mathematical proofs of statistical validity, and is accessible primarily to statisticians. We therefore first illustrate the principles of HMC using simple models, and contrast it with other MCMC algorithms.

MCMC algorithms sequentially generate posterior samples (i.e., vectors containing a value for each parameter), resulting in a finite number of auto-correlated samples which are used for inference [12]. Many algorithms *transition* between samples by proposing a new sample, based on the current sample and tuning parameters, and then accept it with known

probability. If rejected, the current iteration is the same as the previous one.

For example, the widely-used random-walk Metropolis algorithm [39] typically proposes a multivariate normal sample, centered at the current sample, and uses the proposed to current posterior density ratio to determine the acceptance probability. In this case, all parameters are proposed and updated simultaneously, and the covariance of the proposal distribution is tuned to achieve an optimal acceptance rate [40]. Other algorithms update a single parameter at a time, looping through each within a transition. This is the behavior typically used by BUGS, which uses Gibbs sampling if possible, and alternatives if not.

If an algorithm cannot propose samples in regions of the posterior distant to the current state, then it exhibits random walk behavior: multiple transitions are necessary to move between regions, leading to higher autocorrelation and slow mixing. HMC avoids this inefficient random walk behavior because it can propose values (almost) anywhere in the posterior from anywhere else. It does this using a physical system known as Hamiltonian dynamics.

1.3.1 Hamiltonian dynamics

A Hamiltonian system can be conceptualized as a ball moving about a frictionless surface over time (e.g., imagine a marble inside a large bowl). The ball is affected by gravity and its own momentum: gravity pulls it down while momentum keeps it going in the same direction. A set of differential equations govern the movement of the ball over time (its *path*).

There are some important concepts associated with the ball. The *position* of the ball is its coordinate vector (i.e., where it is on the surface), and associated with each position variable is a *momentum* variable. The *potential energy* is the height of the surface at a given position. The *kinetic energy* is related to the momentum, assumed for now to be the sum of the squared momenta. Because the surface is frictionless, the total energy (potential plus kinetic), known as the *Hamiltonian* (H), remains constant over time. Later we will see that, in the context of MCMC, the position vector corresponds to the model parameters and the potential energy to the negative log of the posterior density.

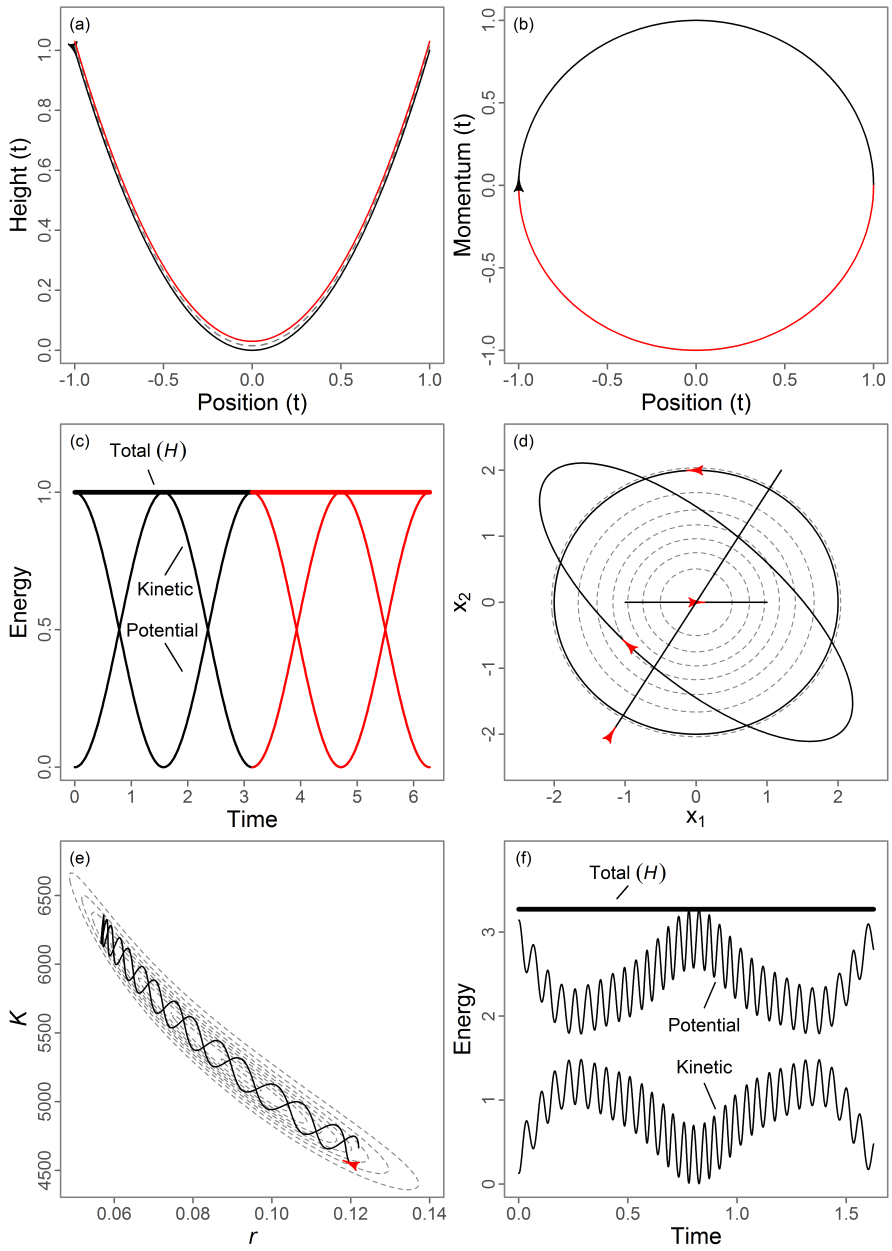

Figure 1.2: Basics of Hamiltonian dynamics. (a) An example where a ball is dropped from the black point, it rolls down the surface over time (t), and momentum carries it up the other side where it reverses direction (red line), returning to where it started. The lines are offset to distinguish black and red paths. The position and momentum variables (b) and energies (c) over time corresponding to the path in (a). (d) Multiple paths for a 2d parabola. Gray dashed lines show posterior contours; initial positions and paths are red arrows and black lines. (e) Partial path (black line) on a posterior of a logistic population model with intrinsic growth rate (r) and carrying capacity (K). Red arrow shows initial position. (f) The energies for the trajectory in (e).

For now, consider the parabola $y = x^2$ (Fig. 1.2a), which has a single position variable (x) and thus a single momentum variable. We place the ball at position $x = -1$ and height (potential energy) $y = 1$, and let it go such that it has no initial momentum or kinetic energy. Gravity pulls it down, building speed over time as potential energy is converted to kinetic energy (Fig. 1.2b,c). Momentum carries it past position $x = 0$, where all potential energy has been converted into kinetic energy. Since there is no friction, it stops exactly at $x = 1$ and $y = 1$, where the potential and kinetic energies return to their initial states (Fig. 1.2c). At this point it will reverse course (Fig. 1.2a-c red lines), and oscillate forever with the energies varying but their sum (H) remaining constant.

Now consider a 2d parabola, $y = x_1^2 + x_2^2$ (i.e., a bowl shape). The position and momentum vectors are of length two, but the kinetic and potential energies are scalars. I place the ball as before, but this time I flick it, imparting momentum with a direction and magnitude (Fig. 1.2d). If flicked sideways, it will move in a circle of constant height. If flicked straight down it will cross the bottom and go up the other side. An elliptical path occurs when flicking the ball at a downward angle. A more complex surface typical of a real model, such as a logistic growth model (see case studies below), leads to more complex paths (Fig. 1.2e,f), but which obey the same principles and intuition as these simple examples.

The principles of Hamiltonian dynamics relate directly to MCMC by providing a way to generate efficient transitions. The ball could move (almost) anywhere given the right length of time and initial momentum, thus providing transitions with directed movement and avoiding inefficient random walk behavior. MCMC algorithms that utilize Hamiltonian dynamics are generally referred to as Hamiltonian Monte Carlo, and I briefly review two: static HMC and NUTS.

1.3.2 Static HMC

Static HMC was the first MCMC algorithm to utilize Hamiltonian dynamics [41]. Although replaced by more advanced algorithms, static HMC is simpler to explain and contains most of the properties relevant for understanding NUTS. A static HMC transition occurs by

simulating the ball from the current position with random momenta for a finite length of time and proposing the state (position) at the end of this simulated, finite path.

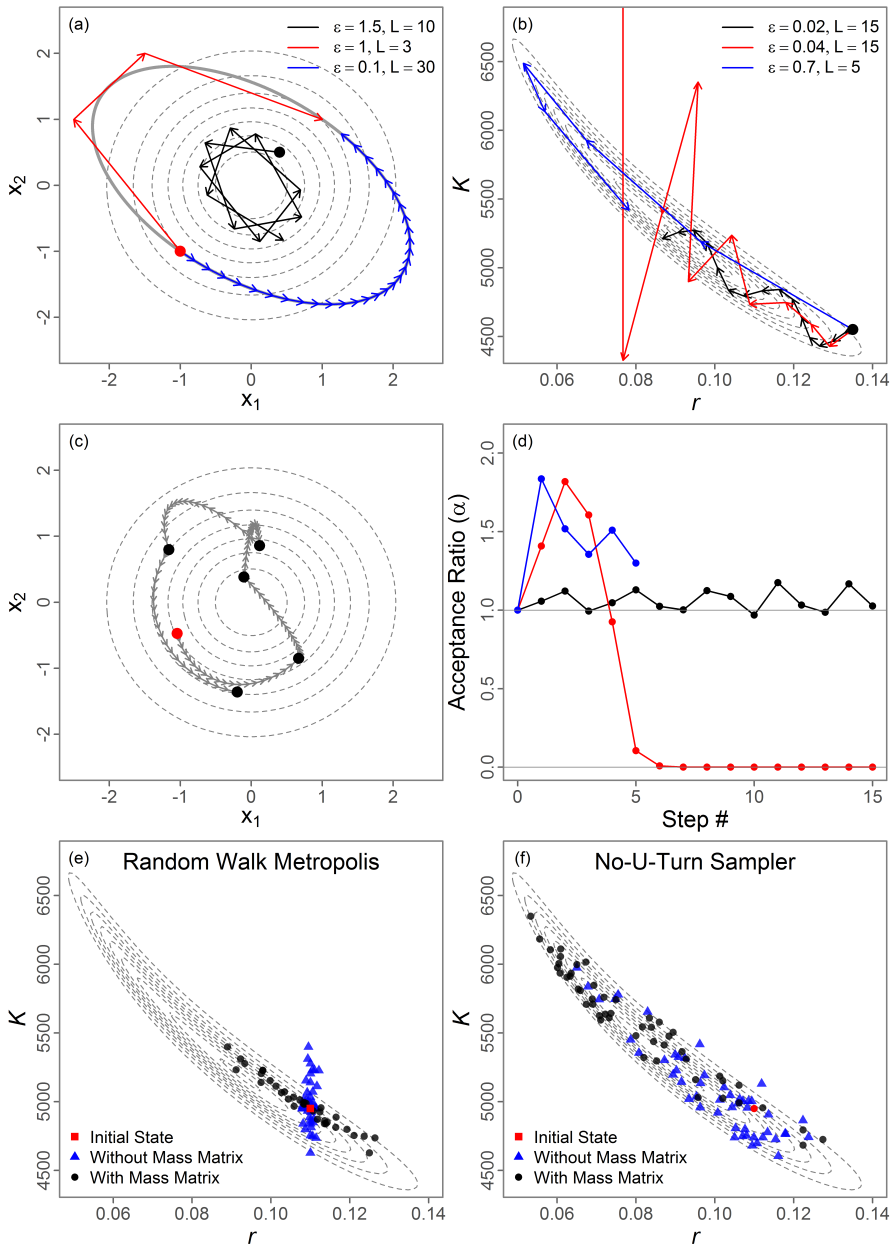

Figure 1.3: Examples demonstrating the basics of HMC. (a) The effect of different step sizes (ϵ) and number of steps (L) on trajectories. The blue and red trajectories approximate the same path (solid gray line), with the same initial position (red point) and trajectory length (ϵL), but opposite momentum. (b) Trajectories on a logistic posterior surface with identical initial position (black point) and momentum vectors. The black trajectory is slow to traverse the surface, while the red trajectory shows accumulating approximation errors, causing it to diverge. The blue trajectory utilizes a mass matrix, making the surface easier to traverse. (c) Multiple iterations of static HMC; black points are and accepted and intermediate steps (gray arrows) are discarded. (d) The acceptance ratios (α) of the trajectories in (b), with corresponding acceptance probability of $\min(1, \alpha)$. Multiple draws from the same initial position using a random walk Metropolis (e) or NUTS (f) algorithm, with and without an appropriate mass matrix (colors).

However, three issues complicate this process. The first is how to simulate movement on arbitrary log-posteriors (i.e., generate paths). Simple models like a parabola have analytical solutions to the underlying differential equations; thus exact, continuous paths are possible. However, for most models the continuous paths must be approximated using a numerical method known as the leapfrog integrator (I refer to approximated paths as *trajectories*). A trajectory depends on the *step size* ϵ and the *number of steps* (L; Fig. 1.3a,b). The position vector at step L is the proposed sample for that transition, while the intermediate steps are discarded (Fig. 1.3c). Approximation errors cause the ball to deviate from the continuous path, and thus H is not constant over time (Fig. 1.3d).

The next challenge is determining the optimal *trajectory length* (i.e., ϵL). If the trajectory length is too short, distant proposals are impossible, leading to an inefficient random walk. If it is too long, the trajectory will retrace its steps (e.g., Fig. 1.3a), which is wasteful computationally. Thus, efficiency depends on the trajectory length, but the optimal length is difficult to determine and a crucial tuning step required for static HMC [42].

The last issue is determining the step size, given a trajectory length. The same length can be attained by taking fewer steps of larger size, or more steps of smaller size (Fig. 1.3a,b). Since each step is computationally costly, the fewer the steps the faster the transition. However, there is a downside to large step sizes: they lead to more variation in H, and in some cases the approximation error accumulates such that the total energy (H) goes to infinity, known as a *divergent transition* (red trajectory, Fig. 1.3b). A Metropolis acceptance step accounts for variation in H by accepting the proposed state with probability $\min(1, \alpha)$, where α is the exponential of the energy lost. Thus, proposals are always accepted if the total energy has decreased, whereas increased energy is accepted with a probability less than 1 (Fig. 1.3d). Increasing the step size reduces runtime, but increases approximation error, leading to more rejected states and divergent transitions, degrading the efficiency of the algorithm. Optimizing the step size is thus another crucial step in static HMC [43]

Given a step size and number of steps, the last step is to specify a kinetic energy function. In HMC it is typically the log density of a multivariate normal random vector where the

covariance matrix is known as the *mass matrix*. Previously I assumed the kinetic energy was the sum of the squared momenta, corresponding to an identity mass matrix. The effect of the mass matrix is to globally transform the posterior to have a simpler geometry for sampling. The variances stretch the posterior so all parameters have the same scale, while the covariances rotate it so they are independent. When successful, the transformed parameters have a scale of 1 and no correlations, resembling iid standard normal random variables (blue trajectory, Fig. 1.3b.)

The mass matrix is analogous to the covariance of the proposal function sometimes used in Metropolis-Hastings samplers, which can have substantial impacts on sampling (Fig. 1.3e). Depending on the model, HMC algorithms can be efficient with an identity mass matrix (Fig. 1.3f), but it will require more leapfrog steps per transition and more time (Fig. 1.3b). Thus to get efficient sampling with HMC, the mass matrix should approximate the covariance of the posterior, but this information is often not known *a priori*.

Specifying an optimal trajectory length, step size, and mass matrix is critical for static HMC to work efficiently, leading it to require expert hands-on tuning and *a priori* knowledge [34]. Fortunately, NUTS automates this process and provides efficient sampling with minimal or no tuning.

1.3.3 The No-U-Turn Sampler

NUTS extends static HMC by automating tuning: neither the step size nor number of steps need be specified by the user. NUTS determines the number of steps via a sophisticated tree building algorithm, which we briefly describe here. A single NUTS trajectory is built by iteratively accumulating steps. In the first iteration a single leapfrog step is taken from the current state so the trajectory has a total of 2 steps. Then 2 more steps are added (total of 4), then 4 more (total of 8), and so forth, with each iteration doubling the length of the trajectory. This doubling procedure repeats until the trajectory turns back on itself and a "U-turn" occurs, or the trajectory diverges (i.e., H goes to infinity). The number of doublings is known as the *tree depth*. The key aspect of this tree building algorithm is that

it automatically creates trajectories that are neither too short nor too long. In practice this means trajectory lengths vary among transitions: it may take 8 steps or 128, depending on the position and momentum vectors.

NUTS determines the step size by adapting it during the warmup (burn-in) phase to a target acceptance rate (`adapt_delta` in Stan). The tuned step size is then used for all sampling iterations. In contrast to static HMC, NUTS does not use a Metropolis acceptance step, so an analogous statistic is used for adaptation. [43] found this target acceptance rate should generally be between 0.6 and 0.9, with larger values being more robust in practice. Thus, NUTS effectively reduces static HMC to a single, user-specified tuning parameter: the target acceptance rate.

1.3.4 HMC in practice

One disadvantage of HMC is that, unlike BUGS, only continuous parameters are possible because discrete parameters do not have gradients. A manual implementation could overcome this by alternating Gibbs updates and HMC [34], and future versions of Stan may implement such a scheme. Alternatively, in some cases they can be marginalized out manually by the user (Chapter 10 and 12, [36]).

Another disadvantage is that HMC is developed using sophisticated mathematics and statistics (e.g., [44]), making it difficult to develop a deep understanding or intuition about their behavior. I provide implementations of the static HMC and NUTS algorithms, written in R [45]. I encourage the interested reader to experiment with the samplers to further their understanding of HMC, while using the faster and more robust Stan implementation for inference of real problems.

NUTS (and static HMC) is similar to other MCMC algorithms: valid inference is conditioned on a converged chain, but this is impossible to prove [12]. The analyst is responsible for assessing convergence before making inference, and for NUTS this includes assessing adaptation. Information about step size, tree depths and mass matrix quantities are reported in the output of a Stan run, and they should be checked routinely. For example,

the adapted step size should be consistent across multiple chains, post-warmup divergences should be minimized (by increasing target acceptance rate) and the maximum tree depth increased if necessary. The user manual [36] has more information, advice on fitting strategies, and details of the adaptation procedure for the mass matrix and step size.

Key concepts that arise when using NUTS in Stan are summarized briefly below:

- Smaller step sizes have higher acceptance rates, but require more steps and thus time. Larger step sizes reject more states and can have more divergences. The optimal step size depends on the model, and is tuned to achieve a *target acceptance rate* set by the user (`adapt_delta`), defaulting to 0.8, but higher values needed for more difficult posteriors.

- The number of steps is determined dynamically for each transition using a tree building algorithm, where the trajectory repeatedly doubles in length until a U-turn occurs. The number of doublings is known as the *tree depth*.

- If the mass matrix approximates the covariance of the posterior, the algorithm 'sees' a simpler surface and is more efficient. By default only the diagonal terms are estimated, accounting for differences in scales, but not correlations, between parameters. Mass matrices with non-zero covariance terms, referred to as *dense*, are available in Stan but are not commonly used.

- The optimal step size depends on the mass matrix, and the mass matrix cannot be well estimated without sampling from the entire posterior, which requires a reasonable step size. Thus sufficiently long warmups are needed for effective adaptation and efficient sampling.

1.4 Case studies

I tested the efficiency of Stan and JAGS for simulated and empirical models from population ecology. To quantify efficiency I used the minimum number of effective samples per unit time,

$E = \hat{N}_{\text{ESS}}/t$, a standard approach to compare among algorithms and software platforms. Further details of how this was calculated can be found in Appendix C. This definition of efficiency (E) can be roughly thought of as the number of independent samples generated per unit time.

I used matching parameterizations for Stan and JAGS, but explored two parameterizations for each hierarchical model and platform. MCMC efficiency for hierarchical models depends on the random effect parameterization, with the *centered* and *noncentered* complementary forms being useful for a broad class of models [46, 47]. Briefly, the centered form models the random effects (τ) directly: $\tau \sim N(\mu, \sigma^2)$, while the noncentered form does it indirectly by letting $\tau = \mu + \sigma Z$, where $Z \sim N(0, 1)$ are the model parameters and implying $\tau \sim N(\mu, \sigma^2)$. See Appendix D for further information and references. I test both forms because the most efficient can depend on the amount of information about σ.

Initial values, random seeds, and length of adaptation can have large impacts, particularly for HMC, so I ran 20 chains of length 40,000 without thinning, initialized from a random sample from a previously run long chain. I used the first half of each chain as a warmup, discarding those samples but including warmup time (but not compilation time) in the total run time. I also did not include time to tune the target acceptance rate for Stan, as the analyst will often determine acceptable tuning parameters during model development. I used default settings for JAGS and Stan, except increasing the target acceptance rate from its default of 0.8 where needed (see Appendix E). I checked convergence, as is typically done for MCMC output, such as the potential scale reduction, $\hat{R}$, being close to 1 [12], in addition to the specific diagnostics for NUTS.

My tests included two simulated models and four models with real data (Table 1.1). The simulated models were a multivariate normal with random covariances (MVND) or repeated correlations (MVNC), both of which were easy to vary in the number of fixed effects and covariance structure. Our simulated non-linear mixed effects somatic Growth model varied in the number of individuals. The first two real-data models were fit to mark-recapture data of birds and differed in their size and complexity: the Redkite model only

Table 1.1: Summary of case studies used to compare efficiency between Stan and JAGS. Further details are available in Appendix E. Latent parameters are those modeled as random effects.

ModelName	**Description**	**Data**	**Parameters (Latent)**	**Hierarchical Structure**	**Reference**
MVND	Multivariate normal with covariances generated from inverse Wishart.	Simulated	Varies: 2-200	None	Simulated
MVNC	Multivariate normal with all off-diagonals set to	Simulated	Varies: 5–50	None.	Simulated
Growth	Non-linear somatic growth with repeated measures	Lengths at age	Varies: 14–406 (10–400)	Normal on growth rate and maximum length, in log space.	Simulated; see [48]
Redkite	Age-dependent survival probabilities	Mark-recapture of birds	5	None.	Section 8.4 of [49]
Swallows	State-space survival and detection with environmental covariates.	Mark-recapture of birds	177 (172)	Year and family effects for survival, family effects for detection.	Section 14.5 of [50]
Logistic	State-space fisheries logistic population dynamics	Annual catch per unit effort; catches	28 (22)	Annual biomass dynamics deviations.	[16]
Wildflower	Binomial generalized linear model of flowering success	Stages, flower, and seed pod production	1101 (1072)	Year effects on intercept; crossed effects on intercept and slope for covariate.	[32]

estimates survival while the Swallows model estimates survival and detection probabilities using environmental covariates in a complex hierarchical state-space formulation. I also fit a state-space Logistic population dynamics model to fisheries data to estimate temporal trends in abundance. Lastly, the Wildflower model was a generalized linear mixed effects model with crossed random effects estimating flowering success. The case studies ranged from 5 to 1101 parameters and were a mixture of hierarchical and non-hierarchical models. Further details can be found in Appendix E, and model files for both Stan and JAGS in Appendix B. I did my analyses using R and the packages `rstan` and `rjags`.

1.5 Results

For the multivariate normal models (MVND and MVNC), the run time of JAGS increased at a faster rate than Stan with increasing number of parameters, although the minimum effective sample size for a given run was similar between the two software platforms. Stan was more efficient by several orders of magnitude because its run time for each sample was faster, and increasingly better with more parameters (Fig. 1.4a,b). For the growth model, Stan consistently outperformed JAGS at higher dimensions for both parameterizations. However, Stan had more variable efficiencies than JAGS with fewer individuals.

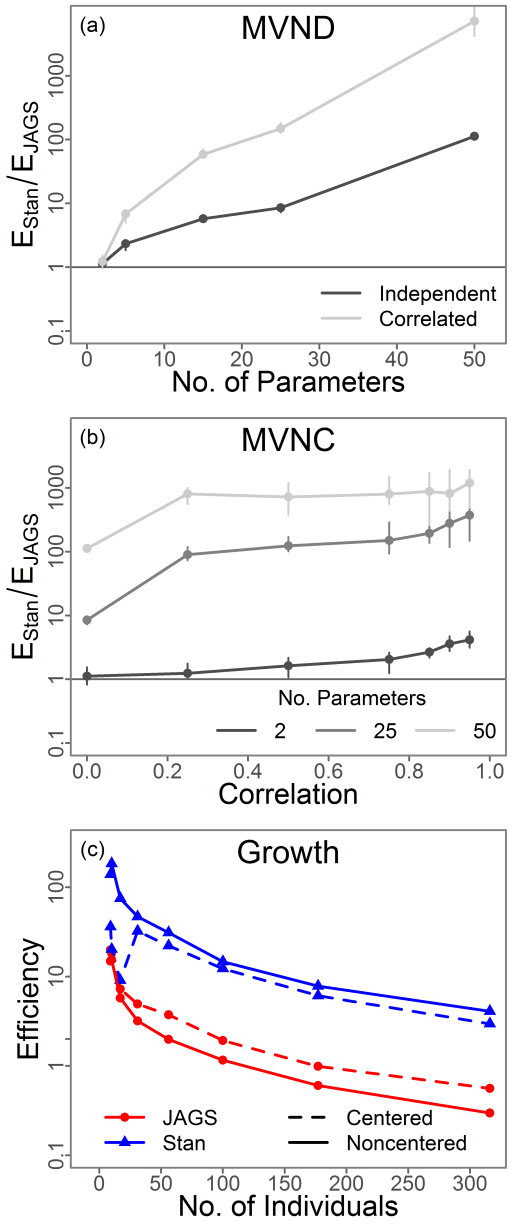

Figure 1.4: Comparison of efficiency (E) for Stan and JAGS across simulated models. The means (points) and ranges (segments) are across 20 replicates. (a) A multivariate normal with increasing dimensionality, either independent or with random correlations from an inverse-Wishart distribution. Ranges are too narrow to be visible. (b) A multivariate normal with repeated correlations on the off-diagonals for varying dimensions (MVNC). (c) A non-linear mixed effects model with two latent parameters per individual (Growth); ranges were left out for visual clarity.

Stan was more efficient for the real-world models as well (Table 1.2), up to 63 times for the Logistic model in the noncentered form. JAGS was faster for the centered Swallows and Wildflower models, but for both the noncentered Stan model was the fastest option overall. Thus Stan was faster for all models (using the optimal parameterization), although the variability in Stan's efficiency tended to be higher than for JAGS (results not shown), likely reflecting HMC's sensitivity to tuning compared to other algorithms.

Table 1.2: Case study results comparing efficiency of Stan and JAGS. Max correlation is the largest absolute pairwise correlation, calculated from converged samples. Efficiency (E) is the number of effective samples per time.

Model	**Random Effects Parameterization**	**Max Correlation**	**Median E_{Stan}**	**Median E_{JAGS}**	**Median E_{Stan}/E_{JAGS} (Range)**
Redkite	NA	0.83	1102.85	302.99	3.54 (1.14–10.03)
Logistic	Centered	0.96	12.35	0.98	12.2 (7.88–34.54)
Logistic	Noncentered	0.96	53.6	0.88	63.33 (18.25–132.02)
Swallows	Centered	0.9	0.12	0.1	0.94 (0–2.96)
Swallows	Noncentered	0.81	0.34	0.1	2.4 (0.1–10.04)
Wildflower	Centered	0.96	0.01	0.06	0.14 (0.02–1.03)
Wildflower	Noncentered	0.96	1.29	0.04	34.2 (13.11–60.7)

I also found clear differences between software platforms in the effect of the parameterization for hierarchical models. For Stan, the noncentered form was consistently faster than the centered form for models with real data: 4.3 times faster for the Logistic, 2.8 times for the Swallows, and 129 times for the Wildflower model. In contrast, JAGS was slower for all three: 0.90, 1.00, and 0.67 respectively. For the simulated Growth model the noncentered form was faster for Stan, but slower for JAGS across all dimensionalities (Fig. 1.4c).

1.6 Discussion

Hamiltonian Monte Carlo (HMC) is a family of MCMC algorithms which utilizes the posterior geometry and properties of Hamiltonian dynamics to make directed MCMC transitions, minimizing the inefficient random walk behavior that degrades the performance for many algorithms used by JAGS. HMC is available to ecologists in the form of Stan, a generic and flexible software package with a similar workflow to JAGS. Here, we demonstrated that Stan outperformed JAGS for all simulated and real-world models from population ecology across a range of dimensions and complexity. Stan was more sensitive to the parameterization of the random effects, suggesting analysts use noncentered parameterizations to improve performance (Appendix D).

My findings corroborate studies from other fields (e.g., [51]), but come with caveats when trying to extrapolate. For example, my simulated models might not reflect nuances in real data, or might not be representative of typical models in other subfields of ecology. Fair comparisons between software are difficult because many factors influence performance, including, but not limited to, priors, tuning parameters, length of chains, and parameterization chosen. For instance, a model that is faster in Stan with a specific prior or parameterization may be faster in JAGS with alternatives. Nevertheless, the results from my case studies suggest that Stan will often be more efficient and thus provide faster inference.

Although my focus was on quantifying sampling efficiencies, the software platforms also behave differently for pathological models. Pathologies are properties of the posterior which obstruct an algorithm's ability to explore the entire posterior, resulting in biased inference of quantities of interest [52]. For instance, posteriors with regions of very low or high curvature (gradients) can be pathological for HMC (section 6.6, [53]). Pathologies affect both Stan and JAGS, but Stan naturally diagnoses them: regions of high curvature are identified by divergences, and flat regions by excessive tree depths [52]. JAGS provides no such feedback, and pathologies may not be apparent using traditional MCMC diagnostics. Pathologies using Stan occur in practice: centered hierarchical models can exhibit biased hypervariances due

to high curvature (Fig. 1.5a). A Stan user can try to eliminate potential bias by reducing the step size, reparameterizing (e.g., noncentering, Fig. 1.5b-d), changing priors, or restructuring their model. Thus, Stan is not only more efficient than JAGS, but it may also provide more robust inference because a user is more likely to detect and eliminate potential biases.

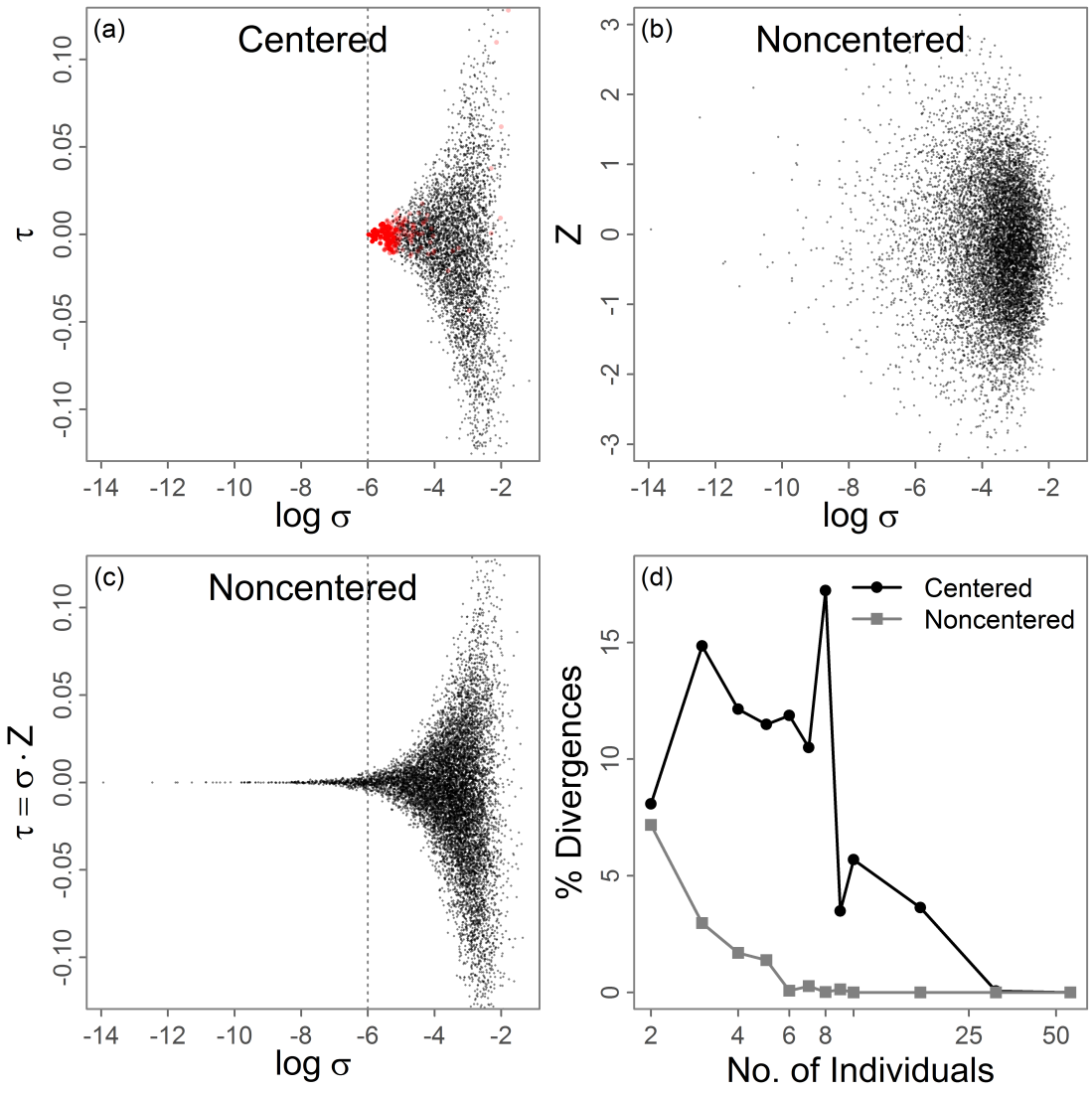

Figure 1.5: Effects of noncentering on divergences and bias for the random effects on growth rate in the Growth model with 10 individuals. τ is the deviation from the mean for an arbitrary individual and the parameters in the centered model, σ its standard deviation, and $Z \sim N(0, 1)$ the parameters in the noncentered model. Samples from: (a) the centered model (target acceptance rate $\delta = 0.95$); (b) the noncentered model ($\delta = 0.80$); and (c) the transformed noncentered parameters, $\tau = \sigma Z$. Divergences in (a), shown in red, arise because the adapted step size is too large for the high gradients at low σ, creating an inaccessible region and leading to biased σ (i.e., no samples below $\log \sigma = -6$). The noncentered parameterization eliminates the curvature, and hence the divergences and bias (c). (d) Median rate of divergent transitions using $\delta = 0.80$ for both parameterizations. As information increases about σ (i.e., more individuals) the marginal distribution of σ narrows, simplifying the geometry and lowering the rate of divergences.

Despite its promise, HMC has some clear disadvantages, with the most critical that discrete parameters are disallowed, such as a discrete latent states or population numbers (e.g., [54]). HMC can still be used if the parameters can be marginalized out analytically, as in the binary states of the Swallows model, and this technique is often possible and can also make substantial improvements for JAGS as well (results not shown). HMC is also sensitive to tuning, despite the automation provided by NUTS. For instance, if warmup periods are too short to effectively explore the entire posterior, then the step size and mass matrix will be suboptimal and efficiency may suffer. Users must also be more involved in assessing tuning for Stan, and be familiar with the principles of HMC to understand diagnostic output from Stan.

There are other HMC algorithms in addition to NUTS, and other gradient-based algorithms for Bayesian inference, which were not tested here. For instance, Riemann Manifold HMC varies the mass matrix along the trajectory [55, 56] and variational inference is a faster alternative to MCMC which approximates the posterior [57]. There are also alternative software platforms not tested here, such as NIMBLE [58] and ensemble sampling [59], and future work comparing these to JAGS and Stan would be worthwhile. Stan is also not the only platform coupling automatic differentiation and HMC that is used by ecologists. Both AD Model Builder [7] and Template Model Builder [60] have HMC capabilities, but neither are as well-developed or mature as Stan (author CCM is a developer of them). My results suggest improving HMC capabilities in these software programs would be worthwhile for their user bases.

The preferred software depends on the situation (Table 1.3), and JAGS will clearly remain a valuable tool when runtime is not prohibitive, but also likely in additional cases such as prototyping models or introducing Bayesian techniques. Stan is clearly the best option for highly parameterized models or smaller models with more difficult geometries (e.g., high or anisotropic correlations). One promising application for HMC is fisheries stock assessment models, which are often extremely large, non-linear hierarchical models that rarely use Bayesian inference because of prohibitively slow run times (e.g., [61]). Many other fields

likely have similar examples where Bayesian inference is currently infeasible, and I anticipate that HMC will make some of these problems tractable for the first time.

Table 1.3: Summary of key differences between JAGS and Stan.

	JAGS	**Stan**
Inference	Bayesian only (MCMC)	Bayesian (MCMC with NUTS & variational inference) and penalized maximum likelihood
Tuning	Automatic with no options	Automatic with options for target acceptance rate (adapt_delta), mass matrix (diagonal or dense).
Discrete Parameters	Use directly	Incompatible - must be marginalized out analytically.
General pros	Easy to use, no tuning, discrete parameters	Scales well with dimensionality, posterior complexity; suitable for hierarchical models, especially the noncentered form
General cons	Few alternatives to reduce runtime when prohibitively slow	No discrete parameters, more difficult modeling language and additional MCMC diagnostics to check
Potential pathologies	No feedback	Diverges and excessive tree depths warn of steep or flat curvature, respectively.

Increasingly large and complex data sets, and powerful software tools, allow analysts to investigate ecological processes which were previously infeasible. Here I demonstrated that Stan, which implements HMC in a flexible modeling platform, is a promising tool when status quo methods such as JAGS are prohibitively slow. I believe Stan should be in the methodological toolbox for every quantitative ecologist since it will extend the boundaries of feasible models for applied problems, and lead to better understanding of ecological processes.

1.7 Acknowledgements

I thank Bob Carpenter and Michael Betancourt for insights on a variety of conceptual issues and constructive feedback on an earlier draft. Margaret Siple, Eric Buhle, Kevin See, Jim Hastie, and two anonymous reviewers provided valuable feedback on an earlier version of this chapter. This publication is partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA10OAR4320148 (2010-2015) and NA15OAR4320063 (2015-2020), Contribution No. 2016-01-23. This work was partially funded in part by a grant from Washington Sea Grant, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award No. NA14OAR4170078. Trevor Branch was also funded by the Richard C. and Lois M. Worthington Endowed Professorship in Fisheries Management. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

Chapter 2

BAYESIAN INTEGRATION IN FISHERIES STOCK ASSESSMENTS: CONFRONTING LONG RUN TIMES

2.1 Abstract

Bayesian inference is widely seen as an appealing alternative to maximum likelihood estimation, but for large, integrated fisheries stock assessments estimation can be prohibitively long. Many factors can contribute to long run times: high dimensionality, model complexity, inefficient algorithms, and geometry of the posterior. Here, we identify causes of slow mixing for six US West Coast assessments written in AD Model Builder and designed for maximum likelihood estimation (some in Stock Synthesis, some custom-built length-structured). We also introduced and tested a new Bayesian algorithm: the no-U-turn sampler. The biggest culprit for long run times was not model size or complexity, but rather parameterizations that reduced the efficiency of the algorithms. Adding priors and fixing poorly-informed parameters (i.e., regularizing) in selectivity curves, and certain random effects, increased run time by 25-82 times. We also found the default AD Model Builder algorithm produced biased estimates in some cases, where the no-U-turn sampler did not, suggesting this new algorithm should be preferred. Thus, regularization by an expert analyst is an instrumental and necessary part of converting a frequentist assessment into a model suitable for Bayesian inference. Between regularization and new software, we expect to achieve run times 10-1000 times faster for most current stock assessment models, opening the door to wider usage of Bayesian methods for management of fish stocks.

2.2 Introduction

Fisheries stock assessments are population dynamics models used to manage fish populations by exploring management actions (e.g., harvest levels) and their effect on the stock [1]. Historically, stock assessments often used production models that used data on catches and relative trends in abundance, but aggregated differences among age, length and sex, and ignored potential temporal changes in biology or the fishery [2]. As data collection increased in quantity and variety, and computation potential increased, stock assessments naturally evolved to incorporate more biological realism. The result was integrated models: complex, highly-parameterized models which incorporate multiple data sources [4, 5]. Integrated assessments are also widely used in bird and mammal management context [62]. Many fisheries stock assessments performed in the U.S. are statistical age- and length-structured models written in the programming framework AD Model Builder (ADMB; [7]). ADMB enables fast and accurate estimates of non-linear statistical models with dozens or hundreds of parameters in either the Bayesian or frequentist statistical paradigms. Such models may be either purpose-built for a particular stock (e.g, [63, 64]), or based on a general model, such as Stock Synthesis [8], which can be applied to many stocks.

Varied data availability, fish life histories, and fishery properties and histories, have driven a wide range of approaches for modeling fish dynamics. Perhaps the most important distinction in integrated models is whether they are fundamentally age structured, or are structured by class or size. The latter is common when ageing is difficult, such as for crustaceans and mollusks (see [65] and references therein).

In these integrated stock assessment models, fisheries selectivity is often the component using the most parameters. Selectivity is the relative probability that a fish of age a or length l from the population is captured (standardized to have a value of one for the most-selected age or length), and is a complex process, since it incorporates both availability and contact (retention) selection [66]. Notably, shifts in effort or gear over both space and time can easily lead to complex shapes that can increase and then decline (i.e., 'dome-shaped') with

increasing age (e.g., [67]). Accurately accounting for selectivity is important [68], and as a result a wide variety of shapes are used, including flexible parametric shapes such as the 'double-normal' in Stock Synthesis, and non-parametric forms such as smoothing splines, and semi-parametric forms, although these are less common (e.g., [69, 70]). In addition, models can allow selectivity to vary over time, either in blocks of years or by allowing a constrained random walk over time in the relevant parameters (e.g., [71]). Given the large number of parameters involved in selectivity curves, over-parameterization is always a concern.

In addition to selectivity, many other processes vary with time, including somatic growth, recruitment, and natural mortality. This time-variation can have important consequences for management quantities [72], but accurately characterizing and estimating this time-variation remains a challenge and often is ignored (e.g., [73]). A common approach to estimate time-variation is to estimate annual deviations away from an average process [74]. These deviations can be modeled as independent, grouped together into time blocks, regressed against an environmental covariate, or auto-correlated as in a random-walk process (e.g., [8]). Most commonly, annual recruitment deviations around the stock recruit curve are modeled as independent deviations, but time-varying quantities like catchability and spatial recruitment apportionment are also used in assessments [75, 76]. Incorporating time-varying features more closely reflects the reality of the biology and fishery, but often substantially increases the complexity and number of parameters of the model, or require subjective assumptions.

In many cases, these deviations are modeled as random effects, meaning they follow a probability distribution whose mean and variance are fixed effects [21]. These mixed-effects (i.e., hierarchical) models are used widely in fisheries science, including stock assessments [22]. Estimating the variance term, in the frequentist paradigm, requires integrating across the random effects to maximize the marginal likelihood. This capability is lacking in ADMB, but two work-arounds exist: use the random effects version of ADMB (called ADMB-RE; [77, 7]), which requires modifying assessments, or approximating the integral externally using an iterative approach, although this is only reliable when there is only a single variance term [74]. Consequently, random effect variances are typically not estimated from the data but

instead fixed at values *a priori*.

As integrated models have become more complex, more emphasis has also been placed on accurately quantifying uncertainty to better inform managers of the risks of alternative decisions [78]. At present, integrated assessments in ADMB predominantly use frequentist approaches to estimate uncertainty: parameters are estimated using maximum likelihood, confidence intervals are obtained under asymptotic approximations, and the delta method is used for derived (e.g., management) quantities [79]. The key assumption underlying this approach is that the maximum likelihood estimator is a multivariate normal distribution with a mean of the maximum likelihood estimate (MLE), and a covariance matrix obtained by inverting the Hessian (matrix of second derivatives of the log-likelihood with respect to parameters) at the MLE [80, 7]. When the likelihood surface violates these assumptions, the resulting inference will be affected. For example, many stock assessment quantities are naturally bounded below by zero (e.g., spawning biomass and depletion) and assuming a symmetric distribution can lead to biased management advice [61]. Likewise, the asymptotic assumption is invalid when the MLE of a parameter is on the boundary of its domain (e.g., a probability must be between 0 and 1). An alternative frequentist method which does not assume symmetry is the likelihood profile, whereby the model is re-optimized at successive fixed values of a single parameter to determine a confidence interval [81]. Likelihood profiles are mainly used in integrated models for identifying data conflicts [82], and I do not further explore them here. Thus, maximum likelihood is fast and reliable in many cases, but it make strong assumptions about the shape of the likelihood surface which may be violated, affecting the reliability of inference.

Bayesian inference is an alternative paradigm to frequentist methods for fitting statistical models, and has a long history in fisheries science, including stock assessment [27]. Bayesian inference uses the same likelihood function as frequentist methods, but combines this with prior information to form the posterior probability distribution of the parameters, given the data and model structure [12]. There are many philosophical and technical differences between these paradigms, and is a continuing debate in the scientific literature (e.g., [83]).

Here, I focus on differences in how estimates and uncertainty are determined. For Bayesian inference, the asymptotic normal assumption is not made and instead probability statements are made by integrating across the posterior. This integration is typically done with Markov chain Monte Carlo (MCMC) algorithms, which generate autocorrelated samples from the posterior distribution [30]. From these samples, approximate integrals can be calculated for quantities of interest, and used to estimate posterior means or credible intervals for parameters and derived quantities.

For ADMB in particular, the default Bayesian workflow is to run a single chain, initialized at the mode, for many (millions) of iterations, discarding the initial warmup period (or burn-in; during which samples are not valid), and then saving every N^{th} during the post-warmup (sampling) period, resulting in 1000 or more samples from the posterior [84]. The resulting samples are checked for signs of non-convergence and used for inference where appropriate [12]. Due to autocorrelation in samples, the effective sample size (ESS) is typically lower than the nominal number of samples and represents the number of independent samples for each model parameter (e.g., C; [85]). As such, it is usually possible to take an ADMB model formulated in the frequentist paradigm (i.e., with no explicit priors) and "flip the switch" to get Bayesian inference using default MCMC behavior and implicit uniform priors.

Due in part to the ease with which the two paradigms can be compared on the same model in ADMB, the advantages and disadvantages for stock assessment models have been explored in previous studies. Notably, simulations showed that Bayesian uncertainty estimates were more reliable, and recommended it as the default method for quantifying uncertainty in stock assessments [79]. Another study compared maximum likelihood and Bayesian uncertainty estimates for two assessments, found that differences between the two methods would lead to different management action, and argued further exploration is warranted when estimates differ [61]. More generally, the Bayesian approach offers a statistically formal way of incorporating prior information from previous studies or expert opinion, and is a natural framework for estimating probabilities of hypotheses and performing decision analyses [81, 27, 28]. Thus, Bayesian methods provide a useful tool for assessment scientists.

However, there are some disadvantages to Bayesian methods. One common challenge is how to specify prior distributions for parameters where nothing is known *a priori* (termed uninformative or vague priors). This is further complicated when parameter transformations are used, such as for unfished recruitment R_0 where a uniform prior in log space may be quite informative in natural space [27, 86], or a non-intuitive formulation of a flexible selectivity pattern. These issues arise in other fields and are a key part of ongoing research and debate (e.g., [87, 88]). However, the most important disadvantage to Bayesian methods for stock assessments, from a pragmatic standpoint, is the extremely long run time compared to frequentist methods. For example one model in [61] was run for 27 days on a dedicated computer (I. Stewart, IPHC, pers. com.) but only ended up with 500 effective samples, where thousands of samples would have been preferable for precise estimates of credible intervals. When models take days, weeks or even months to run, it is difficult to effectively explore model sensitivity or evaluate different cases during development or to address comments during meetings that are often requested by managers or panel reviewers [89]. So, although there are compelling reasons to perform Bayesian analyses on stock assessments, the time needed to attain proper convergence is a major obstacle.

Analyses can be slow for many reasons, but the underlying root cause is needing to evaluate the model a large number (often millions) of times to obtain enough effective samples for sufficiently precise inference. Broadly, this inefficiency (i.e., "slow mixing"; here defined as few effective samples per unit time) is combination of three factors: model run time, posterior geometry, and MCMC algorithm efficiency. The first factor, model speed, is a function of available computing power, the programming language used (e.g., ADMB is written in C++ and will run much faster than an equivalent model written in R [45]), and coding efficiency, and is generally not the major bottleneck for overall run times.

The second factor is the geometric shape of the posterior, which increases run time when the model has many parameters and the posterior has a more complex geometric shape. For instance, efficiency degrades when posterior mass is at boundaries, the posterior has flat tails, or there are correlations between parameters that vary over the posterior (e.g.,

the classic banana shape of a Schaefer model; [90]). Although not widely discussed in the fisheries literature, these geometric properties can increase run time by orders of magnitude. In some cases certain "pathological" regions will be so difficult to sample that the algorithm is unable to sufficiently explore them, and the resulting estimators will be biased [91].

The third important factor for run time is the efficiency of the MCMC algorithm used. Until recently, ADMB had two MCMC algorithms available for use. The default is a modified Metropolis Hastings algorithm which I refer to as random-walk Metropolis (RWM; [39, 92]), which is known to suffer from extreme autocorrelation due to its random-walk nature, particularly in higher dimensions. There is also a rarely-used "hybrid" algorithm [84] which is one of the Hamiltonian Monte Carlo (HMC) family of MCMC algorithms [34]. HMC algorithms are efficient in higher dimensions and more effectively mix in the presence of difficult geometries. Importantly, they are also more robust to pathologies because they warn the user in the form of divergences, which occur when simulated trajectories become unstable due to extreme curvature [93, 91]. Despite this promise, ADMB's hybrid, or static HMC algorithm as is it known in the statistics literature, is notoriously difficult to tune and thus not widely used [34]. However, a new and promising extension of static HMC which requires no tuning was recently introduced: the no-U-turn sampler (NUTS; [29]). Although not currently available in ADMB, NUTS is quickly gaining popularity and is a promising alternative algorithm for stock assessments.

In this study, I confront long run times for a set of stock assessments varying in size, complexity, and structure (age-based, length-based or length-and-age combined). I present guidelines for diagnosing geometric properties that lead to slow mixing, and provide information on how to mitigate these issues. I outline how I incorporated the NUTS algorithm into ADMB and explore differences between NUTS and RWM, including their relative efficiencies and ability to avoid bias. Finally, I compare uncertainty in management quantities between frequentist and Bayesian paradigms, and provide advice on how to overcome common pathologies in Bayesian stock assessment models so that they converge faster.

2.3 Methods

2.3.1 Approaches for minimizing runtime

Broadly speaking, runtime can be reduced by lowering either the number of evaluations, or the time per evaluation. The obvious way to reduce time per evaluation is to increase CPU speed with a new machine. In addition, the model can be restructured to simplify the internal calculations required for a single evaluation. In ADMB, for example, it is possible to separate the model estimation part from forecasts into the future, by placing the code in the "mceval" phase and running projections only on saved samples, as opposed to every iteration of the MCMC chain [84]. Speed increases of up to 50% can also be obtained by increasing length bin widths or lowering the maximum age in a model, thus reducing the size of population matrices that must be calculated in each year [94]. This may be a good alternative particularly during model development, since the posterior will change little but converge faster.

A more general way of lowering run time is to reduce the number of model evaluations. One approach to doing this is improving algorithm efficiency. Here I define efficiency as the minimum number of effective samples (ESS) per unit time across parameters (i.e., the slowest mixing parameter). For a given posterior in ADMB, the analyst can try to increase the efficiency for RWM in several ways, such as optimizing the acceptance rate with respect to changes in step size (i.e., the percentage of iterations which were accepted; [40]). There are also options unique to ADMB to modify the covariance matrix used to generate proposed samples from the current state [84], such as using a mixture of Gaussian and Cauchy distributions (which provides fat tails) for proposals (the mcprobe option), or reducing the correlations among parameters in the proposal matrix (mcrb option). Algorithm efficiency also includes alternatives to RWM, notably the NUTS algorithm, which has strong theoretical justification for being more efficient [44, 47, 52] , and has been used in diverse fields through an implementation in the software package Stan [36]. Since the NUTS sampler was not implemented in ADMB, it was necessary to add it (see below).

The number of model evaluations required can also be reduced through statistical regularization, which, generally, is the use of an external regulator to constrain an optimization problem [31]. For Bayesian inference, this means incorporating informative priors or assuming parameters are fixed. Regularization is necessary because many stock assessments will not have explicit priors if the models were designed for maximum likelihood estimation. In addition to incorporating prior information into the inference, this process often results in much improved MCMC efficiency, so that more informative priors are often recommended to speed up sampler efficiency [36]. Efficiency is improved because many geometric pathologies that degrade efficiency often occur in regions which, *a priori*, would not be expected, and thus can be eliminated in the posterior when appropriate priors are used.

Another case where regularization is important is where a model is over-parameterized, as is often the case for the six-parameter double-normal selectivity pattern in Stock Synthesis. This curve is flexible enough to exhibit both asymptotic (e.g., logistic) and a variety of domed shapes [95], and is used widely in assessments. However, the way the curve is parameterized, and the defaults used by most analysts, present some challenges in the context of Bayesian integration. For instance, if the curve is asymptotic, the parameter controlling the top of the descending limb will have virtually no effect on the selectivity curve. Another issue is that these three parameters are logistically transformed internally to the model (i.e., converted to be between 0 and 1), so that, for instance, the parameter controlling the selectivity at age 0 has minimal impact across a large portion of its range (typically -15 to 5). Specifically, if this parameter is -15 then the selectivity at age 0 is 3.06E-07, and a value of -5 is 6.69E-03, both of which are essentially zero. Both examples lead to large regions of the parameter space with virtually the same log-density, leading to extremely fat tails which do not negatively affect maximum likelihood, but are often difficult for Bayesian algorithms [96]. Either fixing some parameters or adding priors can eliminate this behavior, simplifying the posterior and improving efficiency. In this example, these parameters could be considered nuisance parameters – a consequence of a parameterization designed for maximum likelihood estimation. However, in some cases, pathologies may reflect real posterior features, and should

not be eliminated intentionally. By judiciously constraining, or regularizing, a posterior, the geometry can be simplified and pathologies eliminated, leading to increased efficiency.

2.3.2 Incorporating the no-U-turn sampler in ADMB

NUTS was developed to be the inference engine behind the generic, flexible software tool Stan [36, 37]. Stan has many similarities with ADMB: it uses a template language to build models which are compiled to C++, applies automatic differentiation to calculate gradients, and can perform both frequentist (penalized maximum likelihood) and Bayesian inference. Using Stan as a guide, I added the NUTS algorithm to the core base code for ADMB, making the NUTS sampler available for all existing ADMB models, which number in the hundreds (Fig. 2.1).

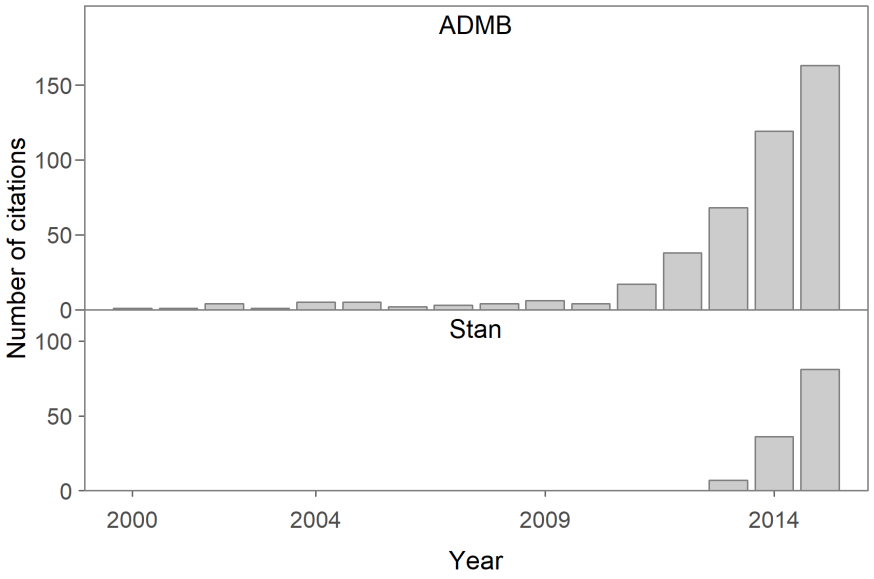

Figure 2.1: Citation patterns between ADMB and Stan across all fields through 2015. The y-axes have different ranges but the same scale.

Stan includes a suite of procedures extending the original NUTS paper, specifically adaptation (either diagonal or dense) of a "mass matrix", which is a matrix whose Cholesky

decomposition is used to (globally) rotate and scale the posterior to make it easier to sample (Fig. 2.2; see section 4.1 of [34]). The RWM algorithm of ADMB also uses this approach so that the mass matrix is relevant for, and has the same effect on, both algorithms. Stan also improves on how a sample is proposed given a single NUTS trajectory by replacing slice sampling with a multinomial draw, an extension known as Exhaustive HMC [42]. Finally, Stan includes an additional HMC algorithm where the mass matrix updates at each step of the trajectory instead of being fixed, known as Riemannian manifold HMC [55, 56]. My implementation in ADMB did not include these more recent improvements, instead focusing on the implementation of NUTS in algorithm 6 of [29]. I did include the ability to use the default mass matrix (i.e., the estimated covariance from inverting the Hessian) or specify a matrix explicitly, typically estimated from a previous run.

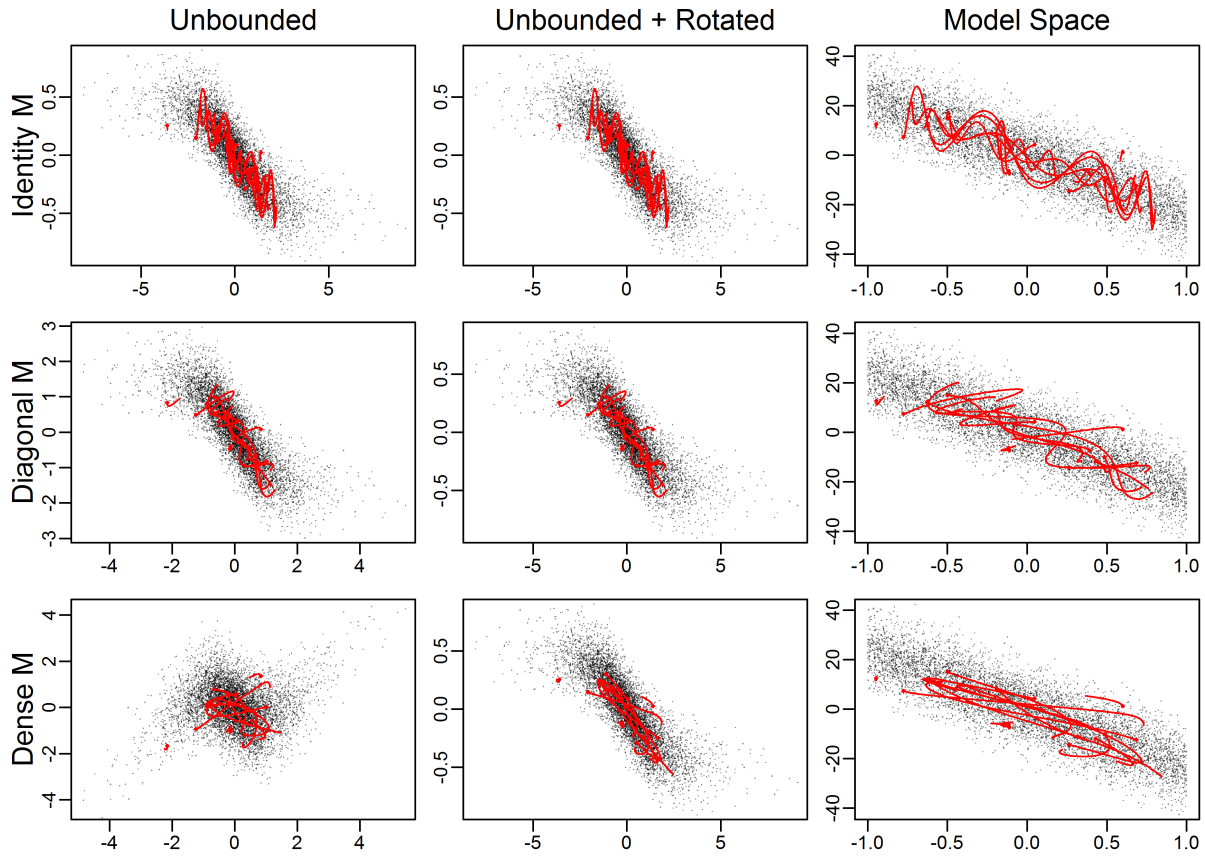

Figure 2.2: The effect of the mass matrix on NUTS trajectories on a correlated bivariate normal model where each parameter is bounded between -1 and 1. Example NUTS trajectories (red lines) are generated from random points with random momenta, with 1000 independent posterior samples shown as black points. Rows indicate which type of mass matrix (M) is used, and columns the three different parameter spaces in ADMB. The algorithm works in the "unbounded" space, and inference is done in the "model space." The dense mass matrix is substantially easier to sample from as evidenced by the straighter trajectories in the bottom right panel. The x- and y-axis are the parameter values in each parameter space.

Accompanying the core functionality Stan is a suite of free, open-source software tools in the software environment R [45], which are used to diagnose and assess the convergence of NUTS chains. For instance, the `rstan` package contains functions to calculate effective sample size, potential scale reduction factor [12], and plotting functions for examining chain behavior [97], while the `bayesplot` package contains functions for analyzing the posterior,

model checking and MCMC diagnostics [98]. Lastly, ShinyStan is an interactive tool for visual and numerical summaries of model parameters, and is particularly useful for examining NUTS chains [99]. I developed an R package, called `adnuts`, to facilitate running MCMC chains (RWM, NUTS, and static HMC) in parallel for ADMB models, and which includes compatibility with ShinyStan and other custom diagnostic tools specific to ADMB fits [100]. This freely-available software package greatly improves the user workflow when conducting Bayesian inference in ADMB.

2.3.3 Case studies

Stock assessments can vary widely in size, complexity, and structure, and I wanted my US West Coast case studies to reflect this. Specifically, I wanted to include age- and length-structured models, custom-built and Stock Synthesis models, as well as models with time-varying components. Further, I choose some models I expected to mix efficiently, and some which I expected to be slow and have issues. Brief summaries of the models are found in Table 2.1; here I outline key characteristics of each.

Cod model: this is a simulated stock assessment based on North Sea cod (*Gadus morhua*), used in simulation tests within the ss3sim framework [101] for a variety of studies (e.g., [73, 94, 105]). This model was chosen because it is quick to run and has stable MLE estimation. The "true" parameters are known in this case because the data are simulated from a known operating model. It is thus an idealized stock assessment model: the sampling process matches the data generating process perfectly (e.g., multinomial composition data and log-normal indices), the true state of nature is time-invariant (except for recruitment deviations), and many key parameters are fixed at the truth. This particular model was modified from [94] to make it run faster by reducing the number of years of data and having fairly wide population length bins.

Hake model: the Pacific hake (*Merluccius productus*) assessment [102] uses Bayesian inference for management (via the RWM algorithm), and has been the subject of a past study comparing frequentist and Bayesian inference [61]. This model converges successfully

Table 2.1: Summary of case studies used. Speed (ms/evaluation) is calculated from random walk Metropolis (RWM) runs in which gradients are not calculated.

Model	**No. parameters**	**Speed**	**Brief description**	**Species and reference**
Cod	61	7.2	Idealized, simulated model, age-structured, in Stock Synthesis	North Sea cod (*Gadus morhua*; [101, 94])
Hake	217	8.9	MCMC results used for management, empirical weight-at-age, in Stock Synthesis	Pacific hake (*Merluccius productus*; [102])
Halibut	195	28.5	Time varying catchability, empirical weight-at-age, Stock Synthesis	Pacific halibut (*Hippoglossus stenolepis*; [76])
Canary	304	199.0	Time-varying growth, three areas without movement, but with different exploitation history, natural mortality varies by age for males, complex selectivity with 31 fleets, recruitment (including apportionment among three areas) varies annually, Stock Synthesis	Canary rockfish (*Sebastes pinniger*; [75])
Tanner	159	90.3	Age and length-structured, custom-built	Bering Sea Tanner crab (*Chionoecetes bairdi*; [103, 104])
Snow crab	334	15.8	Length-structured, custom built	Eastern Bering Sea snow crab (*Chionoecetes opilio*; [63])

using runs of 12 million thinned every 10,000 samples. Individual model evaluations are rapid because the model uses empirical weight-at-age approach and therefore does not need to track length dynamics internally [106, 107].

Halibut model: the coastwide model for Pacific halibut (*Hippoglossus stenolepis*) with short time series (1996-2015), one of an ensemble of four models used in assessing this stock [108, 76]. This model also uses empirical weight-at-age data, has random walks on catchability and selectivity, and estimates early recruitment deviations to determine the initial age structure.

Canary model: the Canary rockfish (*Sebastes pinniger*) assessment [75] includes three areas and has random deviations relating the proportion of recruitment going to each area, implemented as an additive effect in multivariate-logit space [95]. This model estimates 12 selectivity curves, many with the double-normal pattern, and has conditional age-at-length data which causes very slow model evaluations [94].

Snowcrab model: a size-structured model for Eastern Bering Sea snow crab (*Chionoecetes opilio*). This model is custom built for this particular stock [63], and is notable for having the most parameters (n=334).

Tanner model: this model is both age- and size-structured, and is not currently used for management, but rather research for exploring differences between age- and size-structure in the same modeling framework [103, 104]. Here, I use a version customized for Bering Sea Tanner crab (*Chionoecetes bairdi*). One notable property is that the proportion of recruits (P_l) are distributed into initial length bins (l) via a normalized gamma function with parameters M and β controlling the shape as follows:

$$P_l = \frac{l^{M/\beta-1}e^{-l/\beta}}{\sum_l l^{M/\beta-1}e^{-l/\beta}} \tag{2.1}$$

Of these case studies, only the Hake model was specifically optimized for Bayesian inference, and thus some modifications are needed to the other models. The process of deciding how to speed up Bayesian inference on the other models includes many lessons of relevance for others intending to obtain Bayesian inference on their models, as laid out in detail below.

2.3.4 Pilot chains and regularization

The first step in obtaining Bayesian convergence is to run a series of short pilot chains in parallel to assess the geometry of the posterior and detect pathologies that may be slowing mixing. Typically, it is recommended to initialize chains from diffuse points to detect multimodality, but for early pilot chains I started them from the MLE (the ADMB default) to speed up initial exploration. Similarly, the RWM was more stable for these initial runs in the presence of severe issues, and so I used that exclusively during the regularization process. For each case study, I ran 1 million samples, saving every 100^{th}, and using 10% of the chain as a warmup period for adaptation scaling factor for the RWM algorithm. I used pairs plots to examine the slowest mixing parameters (see below), and visually identify the cause. When pathologies existed, I used the following steps to regularize the posterior: (1) fix parameters with their MLE close to a bound at the bound, (2) add stronger priors or fix nuisance parameters which are particularly bad, (3) reparameterize where possible (e.g., if using double-normal and it is asymptotic, convert to a logistic curve). From pilot chains I identified issues, went through these steps, and rechecked the model. I repeated this process until each case study was showing well-behaved geometries, where possible.

2.3.5 Comparing performance between RWM and NUTS

I assessed efficiency for my case studies by running RWM and NUTS chains, with the default mass matrix (i.e., the estimated covariance at the MLE) and an updated one calculated as the empirical covariance of posterior samples from a previous run. NUTS chains were first run for length 500 using a warmup of 10%, which is less than the 50% recommended in Stan because no mass matrix adaptation is done. Each iteration of NUTS has a variable number of evaluations since trajectories are built to be the optimal length, and this depends on the starting point and momenta [29, 93]. Consequently, no thinning is done and run times are also variable among chains. I chose RWM chain lengths to ensure that NUTS runs were similar in time duration, and from these runs I compared the efficiency of the algorithms

and quantified the effect of the mass matrix. I also tested for bias across my case studies, and contrasted estimates of key management quantities between the two statistical paradigms (frequentist and Bayesian).

2.4 Results

2.4.1 Pathologies

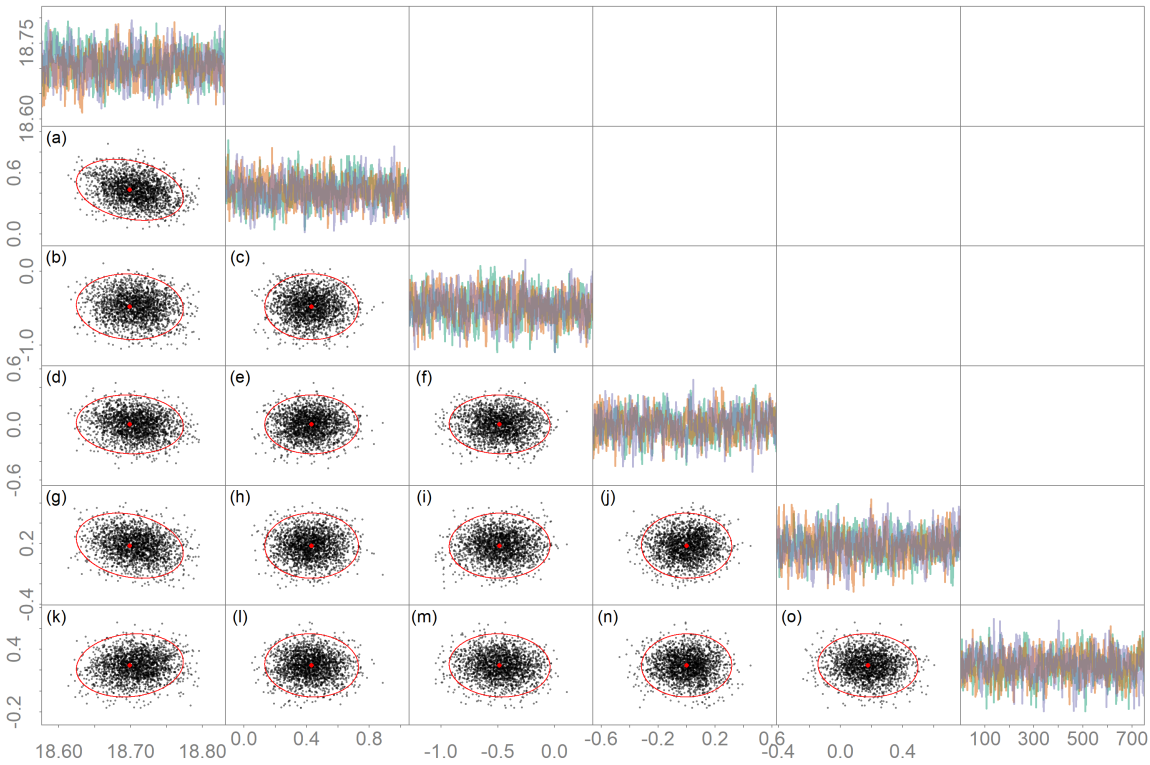

Figure 2.3: Diagnostics for the six slowest mixing **Cod** parameters, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses).

My RWM pilot chains using the default mass matrix demonstrate clear pathologies (regions of the posterior where the geometry challenges the samplers), except in the Cod and Hake

models which mixed well (Fig.2.3 and 2.4). The most common pathology is flat regions of the posterior, such as fat tails. These are regions of the posterior where a wide range of parameter values leads to similar posterior densities. As a result, the samplers occasionally make long sojourns into the tails leading to very slow mixing. One major cause for these is selectivity in the Canary model (Fig. 2.5), which is modeled with double-normal selectivity patterns whose default priors are uniform and wide.

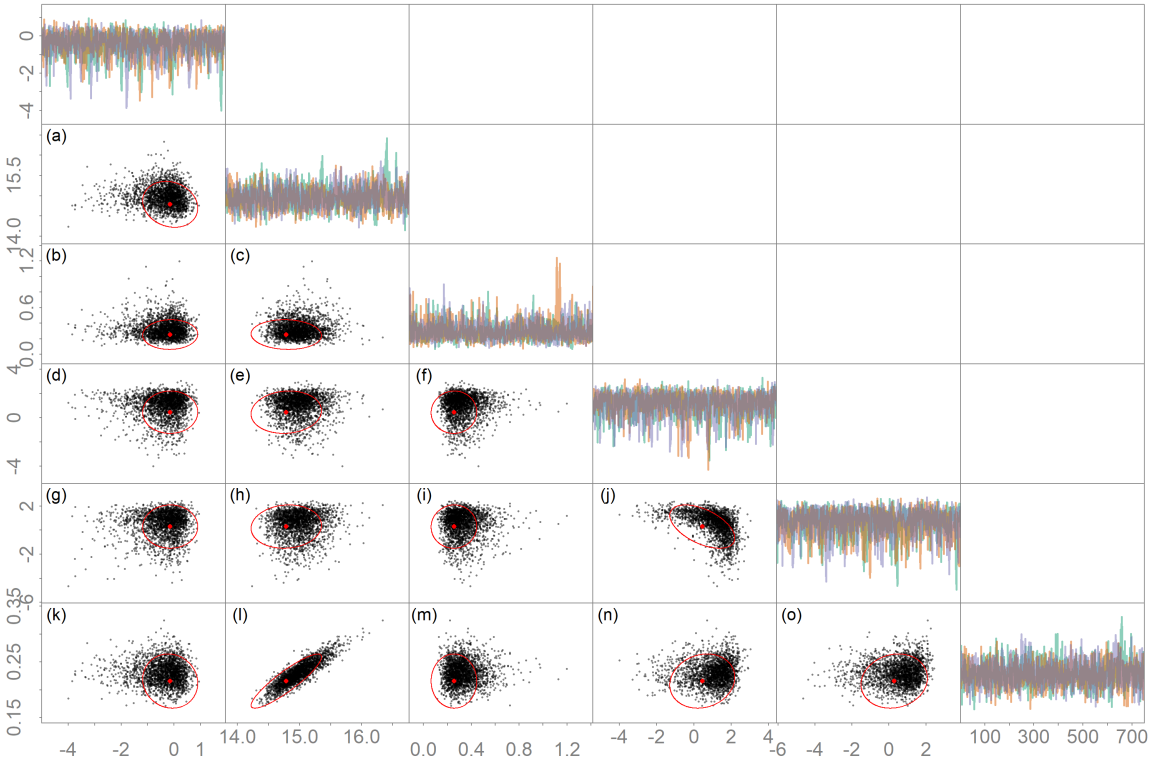

Figure 2.4: Diagnostics for the six slowest mixing **Hake** parameters, from an analysis of five pilot chains with 1000 samples each. Each column Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses).

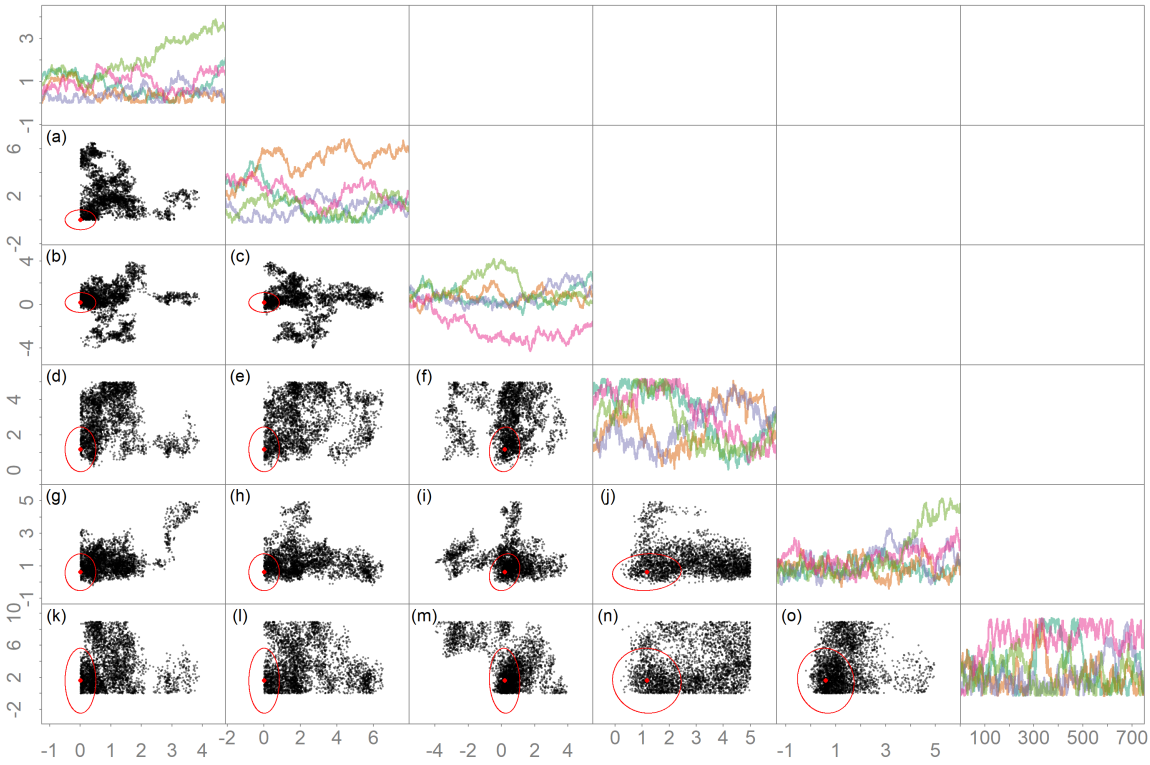

Figure 2.5: Diagnostics for six poorly mixing **Canary** selectivity parameters, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses).

Another common occurrence is parameters being estimated on their bounds (Fig. 2.6), and in these cases, MLE variances were unreliable: sometimes estimated to be either near zero or too big. However, for the RWM chains these parameters were typically not a bottleneck in terms of slow mixing. A surprisingly side note to this result is that the RWM algorithm mixed relatively well despite having a mass matrix based with some components based on inaccurate variances (and covariances) for parameters on bounds. A less extreme, but similar case, is when the MLE was off the bound, but there was large posterior density near the boundary. This can cause computational issues and reduced efficiency with MCMC, and it is thus recommended to reformulate the model to avoid the issue where possible [36].

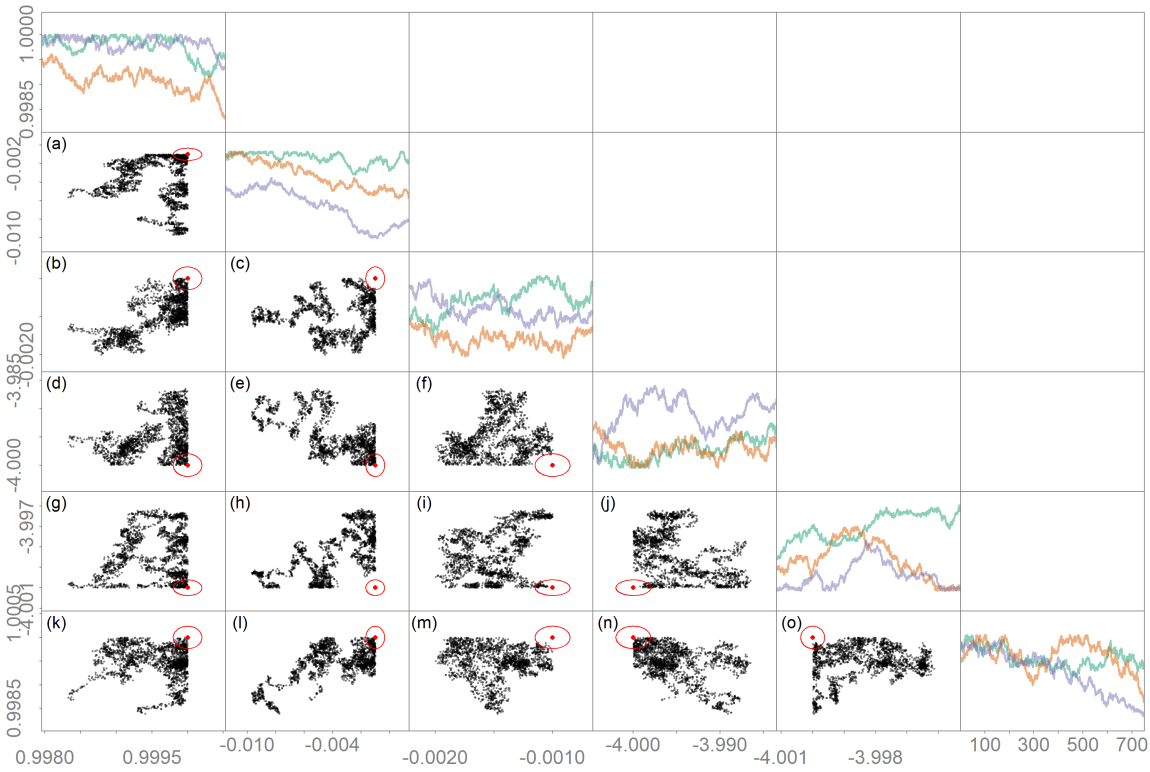

Figure 2.6: Diagnostics for six arbitrary **Snowcrab** parameters which have MLEs near bounds, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses).

Pathologies also included locally-varying correlations between parameters, as typified by early recruitment deviations in the Halibut model (Fig. 2.7). These arise because models often cannot distinguish age classes when the data supporting a large recruitment comes from length distributions. In such cases a large recruitment in year y, and a small one in year $y+1$ can explain more fish of a given length as well as a small one in y and large one in $y+1$. Alternatively, ageing error can lead to difficulty in distinguishing recruitment strength. The core issue is not a single correlation, but that sequential parameters are correlated: y is correlated to $y+1$, $y+1$ to $y+2$, etc. This causes an extreme geometry, when considering all these parameters together (results not shown), and which can substantially challenge

MCMC algorithms and degrade efficiency. Similar correlations were also observed for some pairs of selectivity parameters, but were not as detrimental to mixing as these recruitment parameter correlations in the Halibut model.

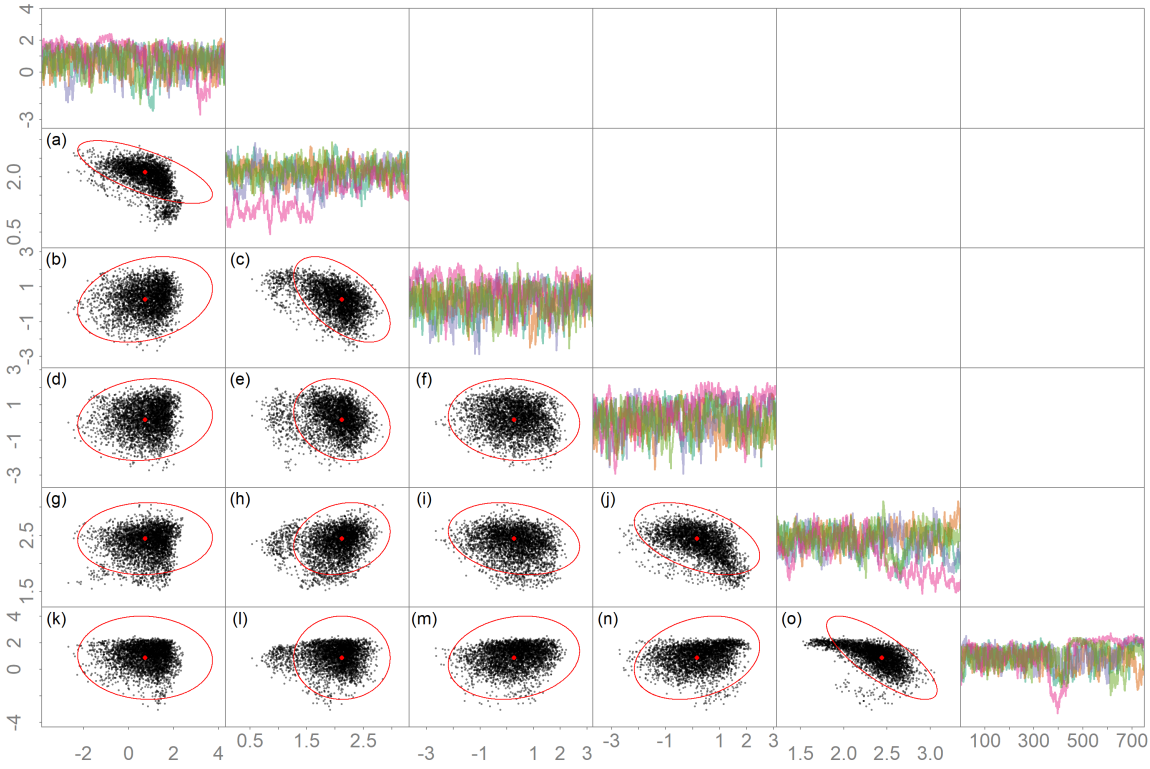

Figure 2.7: Diagnostics for six consecutive early recruitment deviations in the **Halibut** model, from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses).

Lastly, for the Tanner model only I observed bimodality between certain recruitment deviations and a parameter that determines how recruits are allocated to initial length bins, specifically with the M parameter of the normalized gamma function (Fig. 2.8a). The parameter β was estimated at its upper bound of 1.0, and most chains had M varying around $M = 20$. However, one chain escaped this local maximum and explored a region with

higher densit, varying around $M = 30$. Since there is correlation amongst the recruitment deviations, this bimodality affects many parameters throughout the model. Restructuring this model or determining a more biologically defensible prior here is beyond the scope of this study.

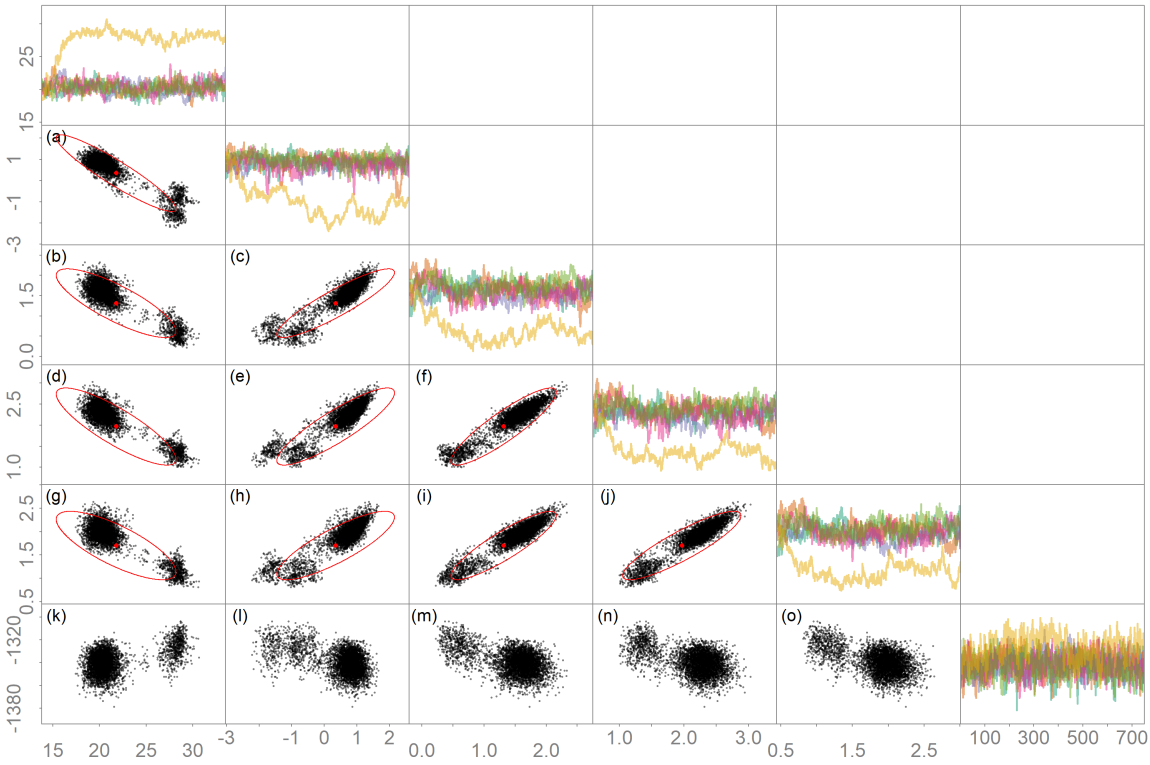

Figure 2.8: Diagnostics for parameters from the **Tanner** model: the M parameter (first row/column), three non-consecutive recruitment deviations, and lastly the log of the posterior density (which has no MLE estimate). These are from an analysis of five pilot chains with 1000 samples each. Each column and row relates to a parameter. The diagonals show traces of the five chains (colors). The lower triangles show scatterplots of posterior samples (black dots; using RWM) and bivariate maximum likelihood estimates and bivariate 95% confidence regions (red dots and ellipses).

2.4.2 Improvements to efficiency due to regularization

I used regularization to constrain each of the models (except Cod and Hake) to try and eliminate pathologies while minimizing the effect on the management quantities. After fol-

lowing the regularization algorithm above, which included fixing certain parameters and adding additional priors (see Table 2.2), the models mixed substantially better (Fig. 2.9). For example, the Halibut model was 82.6 times faster, and the Canary model 25.1 times faster (Table 2.3). I note that ESS, and thus efficiency, are crude approximations and these can be particularly bad when the chains are mixing poorly. Thus, these specific values are likely highly variable, but the trend is clear: regularization of a few parameters can greatly improve mixing for the whole model.

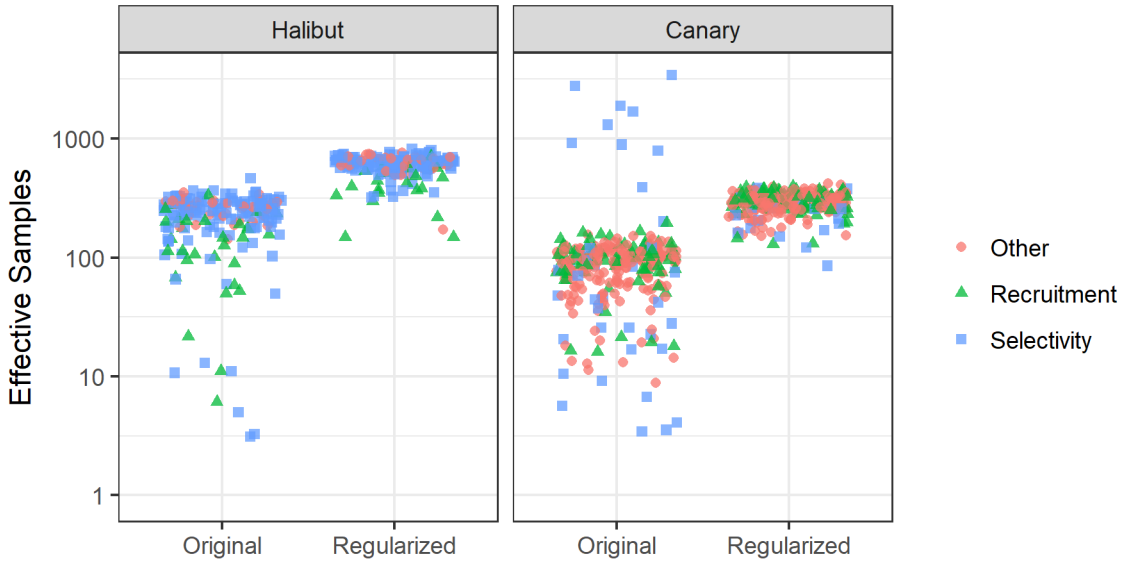

Figure 2.9: Improvement in efficiency due to regularization for two case studies. Each point is associated with a parameter, and the lowest effective sample size is the bottleneck for each model.

Regularization improved mixing, but affected management quantities. For instance, biomass in the Halibut model was estimated approximately 18% lower, but with relatively less change in uncertainty, particularly in 2015 (Table 2.2). The Canary, in contrast, had less than 5% differences in management quantities, but with a large change in uncertainty. Not surprisingly, all estimated standard errors were smaller, as regularization constrains the model by definition. Regularization of the Tanner and Snowcrab models was unsuccessful,

Table 2.2: Regularization and resulting changes to maximum likelihood estimates and standard errors (in parentheses) of select management quantities for the two successfully regularized models. Spawning stock biomass (SSB), depletion (biomass relative to unfished state), overfishing limit (OFL) and maximum sustainable yield (MSY) are common management metrics on the U.S. West Coast.

Model	**Regularization steps**	**Quantity**	**Original**	**Regularized**	**Change**
Canary	Fixed or added priors for 29 selectivity parameters	Depletion (2015)	0.555 (0.049)	0.547 (0.039)	-1.4% (-20.4%)
	and additional variances for indices. Hypervariances	OFL (2015)	1761 (244)	1695 (179)	-3.7% (-26.6%)
	lowered for recruitment and apportionment	MSY	3449 (268)	3411 (241)	-1.1% (-10.1%)
Halibut	15 selectivity parameters fixed or added prior. Recruitment variability reduced from 0.9 to 0.4.	SSB (2000)	467 (32)	550 (29)	17.7% (-9.5%)
		SSB (2010)	182 (16)	216 (15)	18.7% (-3.8%)
		SSB (2015)	190 (19)	221 (18)	16.6% (-2.6%)

Table 2.3: Effective samples for a 12hr (overnight) run using 10 parallel chains using RWM, and the ratio of efficiencies (E) between NUTS and RWM, where applicable.

Model	**Original**	**Regularized**	$E_{\text{NUTS}}/E_{\text{RWM}}$	**Notes and lingering issues**
Cod	70246.6	NA	4.5	None.
Hake	10702.9	NA	0.8	None.
Halibut	98.0	8092.4	2.4	None.
Canary	15.1	378.5	0.4	None.
Tanner	2.8	31.3	NA	Bimodality, parameters near bounds, sensitive to changes so regularized version not compared.
Snowcrab	170.7	174.0	NA	Many parameters near bounds, sensitive to changes, so regularized version not compared.

as both had serious issues remaining. Further exploration would require deeper changes than possible here, and so instead I dropped these two case studies from my speed tests.

2.4.3 RWM vs NUTS and effect of mass matrix

The efficiency differences between algorithms were not consistent, with NUTS being faster for the Cod (4.5 times) and Halibut (2.4 times) models, but not for Hake or Canary (0.81 and 0.4 times, Table 2.3; Fig. 2.10). However, in some cases the NUTS algorithm explored a wider region of the posterior, at the cost of slower mixing, in which case efficiency comparisons are not relevant. One prominent example is recruitment deviations in the Halibut model (Fig. 2.11).

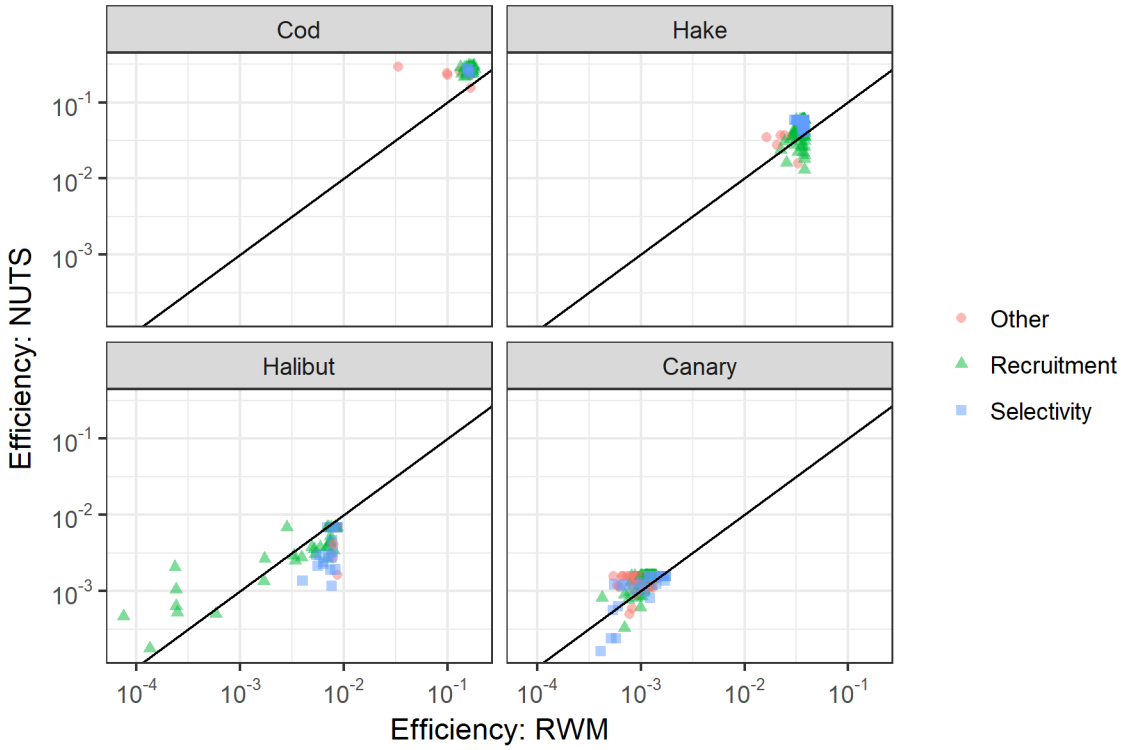

Figure 2.10: Comparing efficiency between algorithms. Scatterplots of efficiency of RWM vs NUTS for each regularized model. Each point is a parameter, and those below the line suggest RWM mixes better, and vice versa for NUTS.

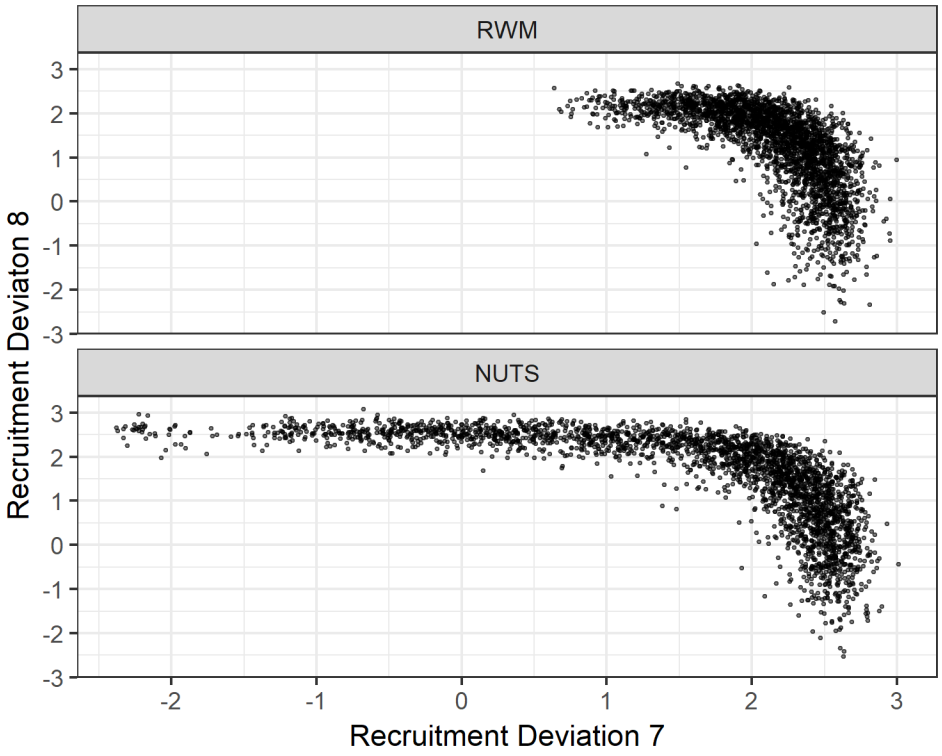

Figure 2.11: Samples from two adjacent recruitment deviations in the regularized Halibut model for the two algorithms. The NUTS algorithm is exploring a larger region of the posterior, suggesting that RWM is biased for this run.

I also found major differences in the sensitivity of the algorithms to the mass matrix used. The RWM sampler was almost always better when the MLE mass matrix was used compared to an updated one estimated from a previous run (Fig. 2.12). In contrast, NUTS mixing was substantially improved with the updated mass matrix for Cod, slightly improved for Hake and Canary, and mixing was similar for Halibut. Based on these models, it is clear that NUTS' efficiency is more sensitive to getting the mass matrix correct than RWM.

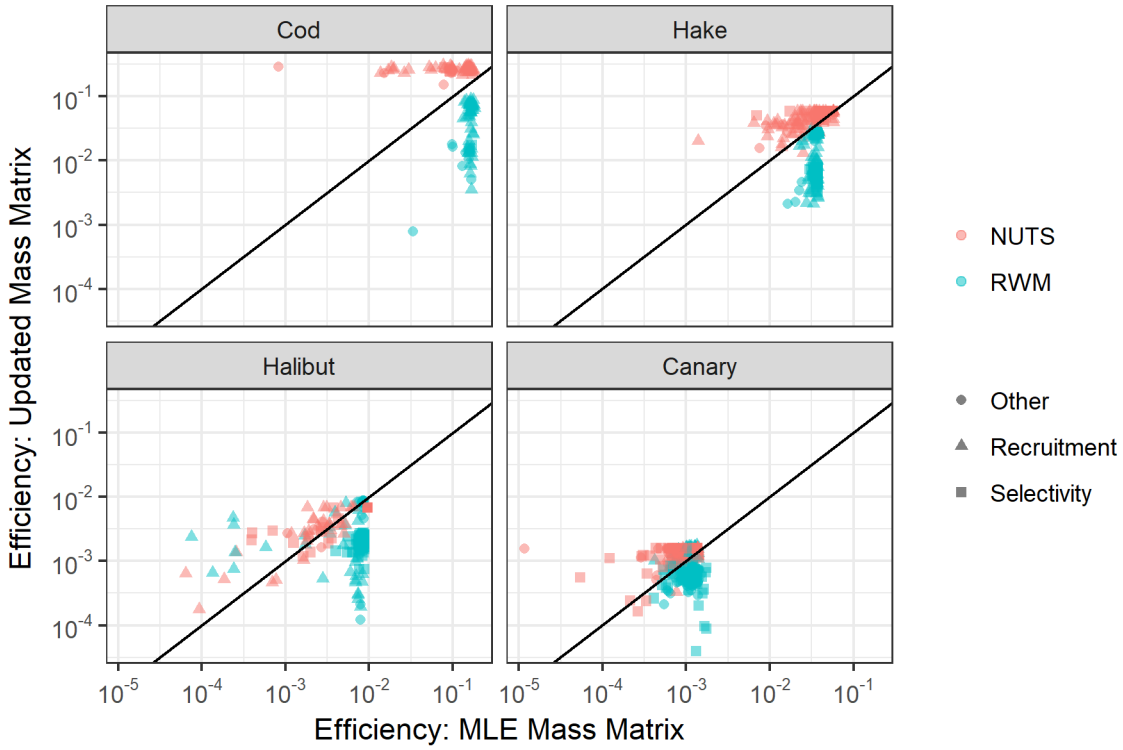

Figure 2.12: Effect of sampling efficiency for both algorithms on regularized models for the default (MLE) mass matrix or an updated "dense" one. Parameters below the line mix better with the MLE mass matrix, while those above are improved by estimating a dense matrix.

2.4.4 Uncertainty estimates

The Cod model was the only one where the estimated parameter variances were consistently equivalent between the frequentist (MLE) and Bayesian paradigms (Fig. 2.13). The Hake model had MLE variance estimates that were consistently smaller than those from marginal Bayesian posteriors, while the other models were more variable, and showed no consistent pattern. Some parameters had variance estimates that differed by more than 50%, particularly selectivity parameters in the Canary model (Fig. 2.5). These were due to MLE estimates being near bounds, which interferes with the asymptotic estimation of parameter uncertainty (i.e., inverse Hessian).

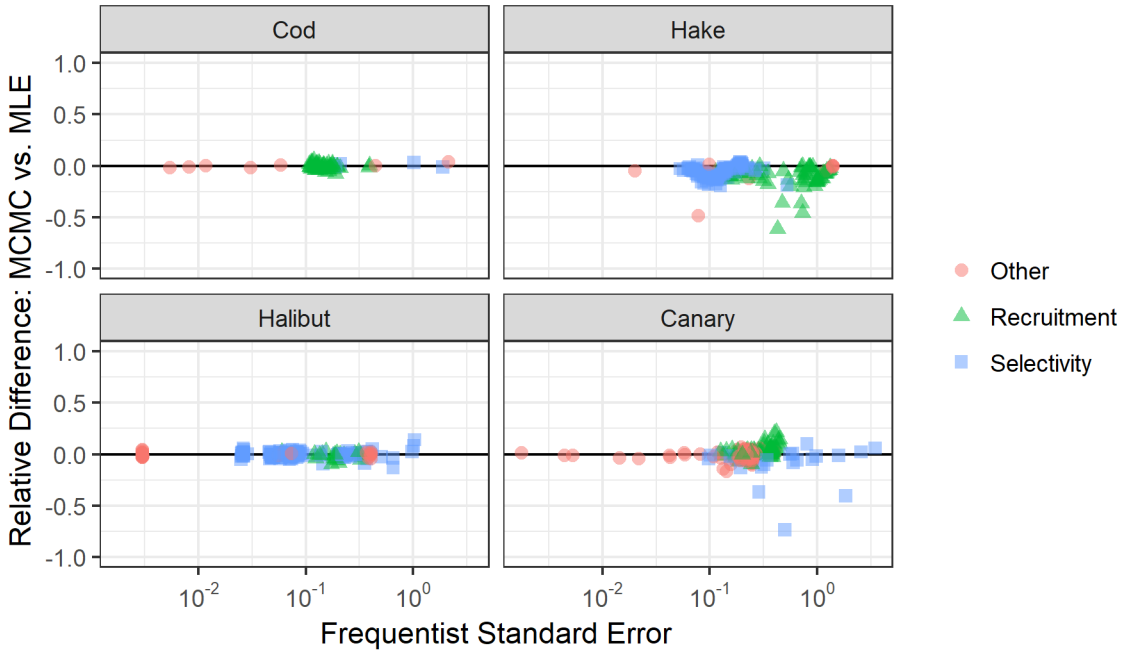

Figure 2.13: Relative differences in parameter uncertainty between the two paradigms after regularizing. Frequentist estimates are standard errors and Bayesian estimates are marginal posterior variances (from RWM). Values below the line indicate that frequentist uncertainties are smaller.

For Hake and Halibut, I found that Bayesian estimates for derived quantities were higher, corroborating a previous studying examining an earlier version of the assessment ([61]; Fig. 2.14). However, the Canary model did not show this pattern, and estimates were generally similar between the two paradigms.

2.5 Discussion

Bayesian analysis is rarely done for integrated stock assessments, despite known advantages, because run times can be several orders of magnitude longer than is practical. Poorly specified models, with respect to numerical integration, were the biggest reason for long run times because this results in a posterior geometry that is challenging and degrades efficiency. Fortunately, judicious regularization (i.e., constraining the model with priors or fixing param-

eters) can improve mixing substantially, sometimes with minimal effect on key management outputs. By combining new software (adnuts), a new algorithm (NUTS), judicious regularization, and ever increasing computational power, we believe reasonable run times are possible even for the largest, slowest fisheries assessment models.

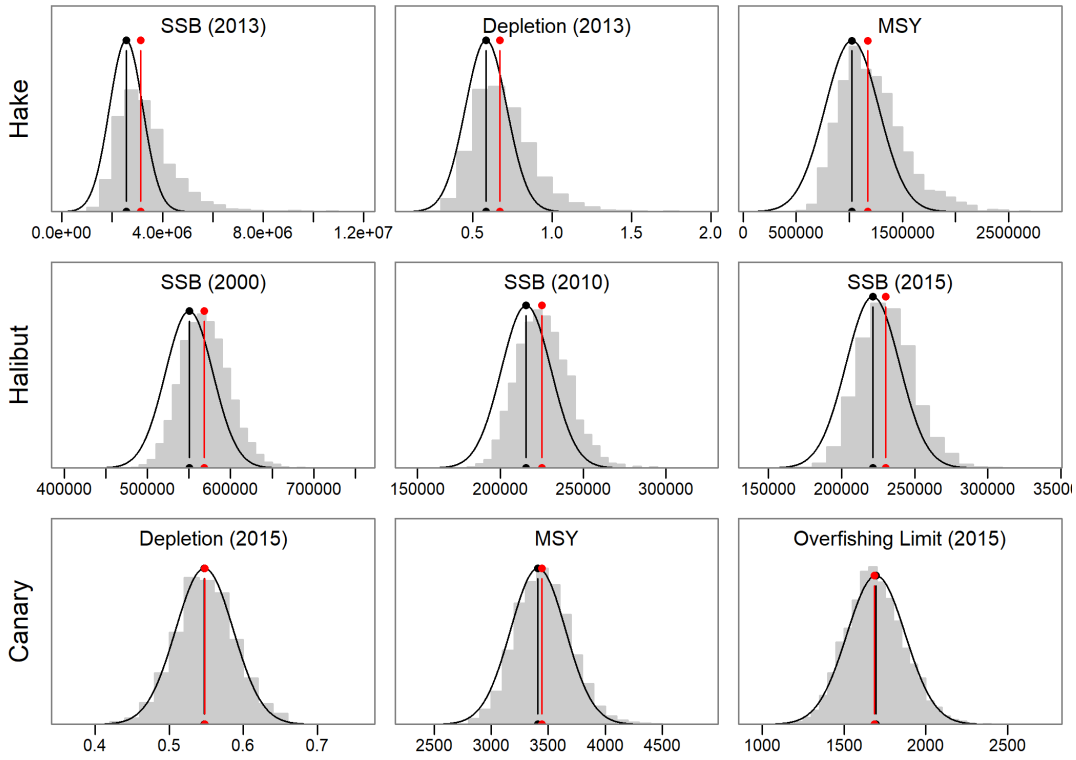

Figure 2.14: Comparison of estimates of key management quantities for 3 regularized models. Posterior distributions are shown as gray histogram, with posterior median as red vertical line. The asymptotic estimate from the delta method, assumed to be normal, is shown as a black curve with the estimate a vertical black line. Different management qualities are relevant among models due to differences in setup. Spawning stock biomass (SSB), depletion (biomass relative to unfished state), and maximum sustainable yield (MSY) are common management metrics on the U.S. West Coast.

2.5.1 Pathologies and improving mixing

All models examined here, except Hake and Cod, displayed poorly defined posteriors in their original state. Selectivity parameters were the most common culprit: posteriors for

these parameters were often stuck at bounds, had fat tails, or displayed non-linear correlations. Perhaps this is not surprising, as my case studies here, and most other fisheries stock assessments, are developed for maximum likelihood estimation in the context of a review process where many versions of the assessment are run. In this context, parameters poorly informed for one configuration may not be for another (e.g., if data weights are changed), so it makes sense to leave estimation of them on. However, we reiterate that maximum likelihood estimation assumes the likelihood surface is multivariate normal, parameters are not on boundaries, and the covariance can be estimated by inverting the Hessian at the MLE, provided the Hessian is positive definite. All the original models could estimate derived quantities with uncertainty reliably, and indeed they were very similar to the Bayesian posteriors. However, it is a dubious practice at best to ignore clear violations of these assumptions just because a model converges and produces a covariance matrix. Occurrence of these pathologies will affect frequentist and Bayesian estimates alike, and I stress that assessment models should meet their underlying statistical assumptions before using them for management. Thus, the following guidelines can improve both Bayesian and frequentist assessment models.

2.5.2 Guidelines for Bayesian integration of new models

Stock assessments are large, complex models and successful Bayesian inference is a difficult task. In my experience, just "flipping the switch" and running a MCMC sampler is insufficient, and needs to be paired with diagnosis of causes for slow mixing. Below I provide guidelines for approaching this task, and note that most models would converge overnight after a few days of work regularizing.

Guidelines:

1. Incorporate all available informative priors on model parameters.

2. Run parallel pilot RWM chains (at least 3) started from parameter MLEs. A good starting place is chains long enough to obtain 1000 samples after thinning every 100,

but a lower thinning rate for slow models may be appropriate.

3. Identify slow mixing parameters using pairs plots from my package adnuts (e.g., Fig. 2.5) and apply appropriate fixes to model: regularization or reparameterizing.

4. Rerun pilot chains and compare differences in frequentist estimates of key management quantities.

5. Repeat steps 2-4 until the model is sufficiently well-defined.

6. Run parallel pilot chains with NUTS, producing 500 samples with no thinning. If divergences (i.e., trajectories that encountered extreme gradients – an effective warning of potential issues) exist, identify the cause. Solutions include: more regularization or reparameterizing, and increasing the target acceptance rate. Check for pathologies and again apply fixes.

7. Run inference chains using NUTS retaining 500 samples, using updated mass matrix estimated from previous NUTS chains.

8. Compare differences in parameter and derived quantity estimates between frequentist and Bayesian inference. If different, check assumptions of frequentist model, and explore alternative priors for key parameters that may be influencing results.

This process is an art form rather than an exact science, but most pathological issues are caused by poorly specified priors and over-parameterized models. That is, pathologies often exist in regions of the posterior which are implausible with prior information and should have negligible posterior density. Some biological parameters may have informative priors from previous studies, but also an analyst usually has prior information on the scale of many parameters, as evidenced by the routine use of bounds in assessment models. In cases like flexible selectivity patterns, the priors are typically uniform, but the implied prior on

selectivity itself is difficult to predict and should be explored visually (Fig. 2.15). In some cases, these pathologies may reflect important posterior features, such as the bimodality in the Tanner model. In such cases, the algorithms will not reliably sample the posterior and it may be better to run two or more versions of the models and combine them together as an ensemble [108]. Regardless, regularization by an expert analyst is an instrumental and necessary part of converting a frequentist assessment into a model suitable for Bayesian inference.

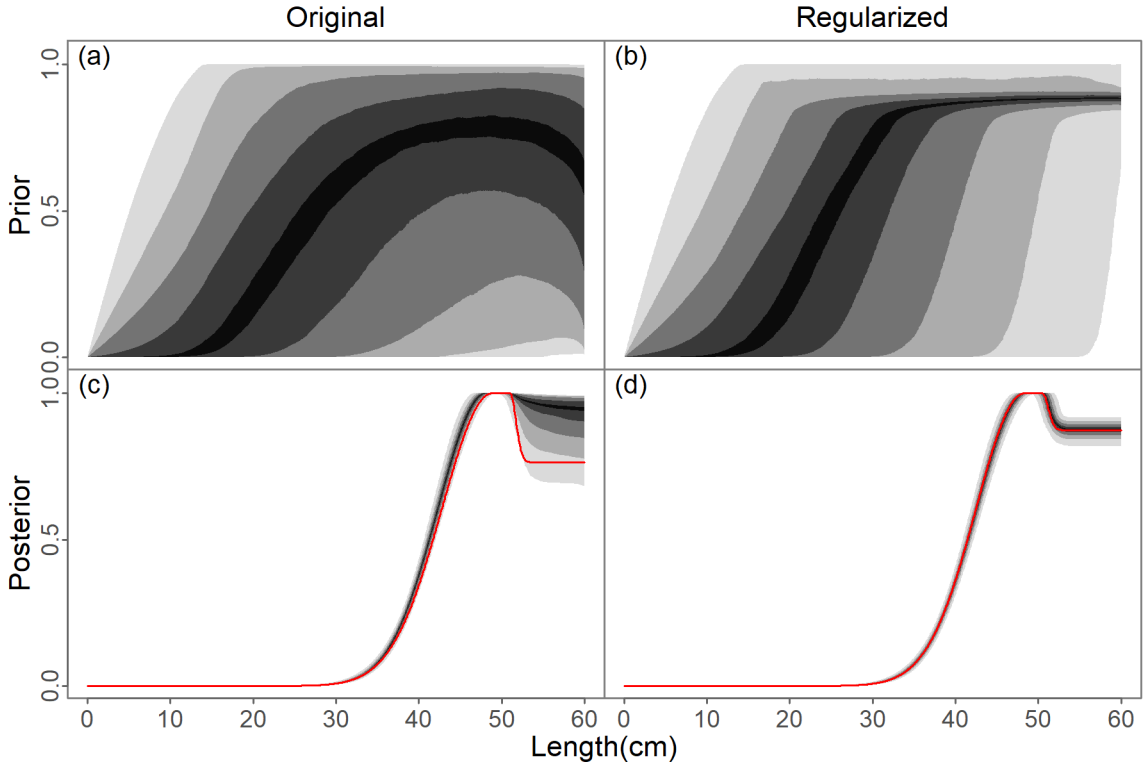

Figure 2.15: The double-normal selectivity pattern for the first fleet in the Canary model for the original with pathologies (see Fig. 2.5) and regularized versions of the model. The implied prior (top row) and posterior (bottom row) are shown as shaded regions, while the red line denotes the MLE.

2.5.3 Advantages of NUTS

The effect of pathologies on mixing, and thus regularization, also depends on the MCMC sampler used. The NUTS algorithm I added to ADMB and applied here promises two advantages over RWM: greater efficiency, and a natural method of diagnosing pathologies by warning of divergences and excessive trajectory lengths [29, 52, 53]. I found that NUTS was faster for the Cod and Halibut models, but not the Hake or Canary models, but note that fair efficiency comparisons between algorithms is difficult due to their stochastic nature and how they deal with pathologies. NUTS better explores more difficult regions (i.e., those caused by pathologies) but this exploration reduces algorithm efficiency, increasing run time. In any cases where RWM is unable to explore the same regions, and is therefore biased, efficiency comparisons are irrelevant. The most clear example of bias comes from the regularized Halibut model (Fig. 2.7). The RWM algorithm did not explore some of the posterior at all, and thus was able to appear more efficient. Without NUTS, it would be impossible to know that part of the posterior was not being sampled, which highlights the value of applying NUTS to the Halibut model. This bias is unlikely to have important management quantities in this case, but there is no guarantee this is the case for other models. As such, I believe NUTS is a valuable tool for analysts performing Bayesian inference on stock assessments in ADMB, and recommend it as the default algorithm for assessment models.

2.5.4 Random effect integration

Random effects are clearly a powerful tool for modeling various aspects of fisheries, both biological (e.g., recruitment and growth) or fisheries (catchability or selectivity). Other software packages, like ADMB-RE [77] and TMB [60], are capable of this estimation and are used for some assessments (e.g., [109, 110]). A side effect of implementing NUTS in ADMB was the addition of the NUTS algorithm for TMB models, and I expect my results for ADMB to also apply to assessments built in TMB [100]. However, in the current framework of ADMB, the variances of these random effects cannot be estimated, and thus is typically fixed

at a subjective value. Although not explored here, I note that if a variance is programmed as an active parameter in ADMB, then it can be estimated because MCMC does the integration. The challenging part is that the ADMB workflow requires an estimate of the global covariance of the posterior, which can cause the inverse Hessian to fail if the mode is at zero. Priors could be used to keep such variances estimable, and then the resulting mass matrix could be used after turning off the prior. Alternatively, mass matrix adaptation could be used, as done in Stan. This may be a useful approach to estimating multiple random effect variances without rewriting assessments in a new software framework [74].

2.5.5 Implications for management

Statistical inference from stock assessments provides the backbone of scientific evidence on which decisions about management are made, and thus it is important to identify differences between frequentist and Bayesian inference [61]. While parameter uncertainty differed between these paradigms (Fig. 2.13), these typically led to small differences in estimates and uncertainties of key management quantities (Fig. 2.14). As argued in [61], differences most likely occur when quantities are skewed and thus a symmetrical normal approximation is inaccurate. My results corroborate this argument, because the quantities in the Canary model were less skewed than in the Hake and Halibut models, and thus the normal approximation more appropriate (Fig. 2.14). In some cases, it may be possible to transform some quantities (e.g., do delta method on log of terminal depletion) to better satisfy the asymptotic normality assumption used in determining maximum likelihood standard errors. In any case, the difference between statistical paradigms is likely to be less than structural differences within a model (e.g., whether to use full time series or not; [108]), or changes in the model over time under different assessment authors [111].

2.5.6 Future Research

I see several avenues for fruitful extensions to the work here. First, it was clear that the double-normal parameterization used in the case studies was particularly challenging for

MCMC samplers. This suggests that it is warranted to explore alternative parameterizations for flexible selectivity patterns that are more commensurate with Bayesian integration in addition to maximum likelihood. One intriguing option is a non-parametric or semi-parametric approach [70]. There are also many avenues of improvement to be made on the algorithm side, as research on HMC algorithms is progressing quickly. For instance, Riemannian manifold HMC adjusts the mass matrix at each step of a trajectory, and thus is better able to handle locally-varying correlations [55, 56]. Variational inference, a faster alternative to MCMC which approximates posteriors, could also be explored (e.g., [57]).

However, likely the most important next step is to explore how well fisheries stock assessment models perform when written in the software program Stan. Models would almost certainly run substantially faster, and alternative algorithms (including Riemannian manifold HMC and variational inference) and future developed and implemented algorithms would immediately be available to use. It would be a daunting task to code these new algorithms into ADMB, but it would also take a concerted effort to rewrite a model like Stock Synthesis in Stan. In any case, an important first step is to try new assessments in Stan to see how much improvement could be expected. I believe that better parameterizations coupled with Stan could result in several orders of magnitude faster Bayesian estimation. Such improvements open the door to exploring extended model features (e.g., spatial or spatiotemporal effects) within the assessments, leading to models which likely would more accurately reflect the stock.

2.5.7 General conclusion

Assessment models are adapted and improved within a management framework of scientific review panels, often requiring updated results overnight. Given its speed advantage it is not surprising that maximum likelihood estimation is the predominant method for inference. In contrast, extremely long Bayesian run times pose a major challenge in this context. Here, I showed that orders of magnitude improvements can be achieved with regularization, faster algorithms, and parallel chains. Some models may still take days to converge with sufficient

samples for inference, and even then, the resulting inference may be essentially the same for key management quantities as frequentist inference (although this is not guaranteed nor known *a priori*). Despite this, I argue that Bayesian inference still provides value for two reasons. First, it helps an analyst diagnose issues with the geometry of the model, highlighting issues with the model that are not apparent otherwise, and which may also affect frequentist inference. Specifically, Bayesian integration can be used to check whether the frequentist asymptotic assumptions are met, leading to more statistically robust models to use for management. Another important role for Bayesian inference is it provides a formal way to incorporate prior information and is a natural way to perform decision analysis [27]. As the debate over statistical paradigms continues, we acknowledge their advantages and disadvantages, and highlight the value in comparing inference between the two methods on the same stock assessment.

2.6 Acknowledgments

I thank Dave Fournier and Johnoel Ancheta for help with the ADMB source code, and Kasper Kristensen and Hans Skaug for helpful discussions on the technical aspects of the TMB source code. I also thank Cody Szuwalski, Caitlin Akselrud, Ian Stewart, and Ian Taylor for help with the assessment models. I also thank Bob Carpenter and Michael Betancourt for helpful discussion on HMC in general, and Bayesian statistics in general. I also thank Jonah Gabry for modifying ShinyStan to be compatible with my software.

Chapter 3

THE EFFECT OF HOOK SPACING ON LONGLINE CATCH RATES: IMPLICATIONS FOR CATCH RATE STANDARDIZATION

3.1 Abstract

Catch per unit effort (CPUE) is widely used as an index of population abundance to inform stock assessments that are used to estimate population status and set fishing policies. For CPUE to be an unbiased index, influences on CPUE that are not related to population abundance (e.g., spatial patterns of effort and changes in gear efficiency) must be accounted for in a CPUE standardization. In longline fisheries, one important factor affecting CPUE may be the spacing between hooks ('spacing effect'), as this is hypothesized to affect the effective effort of a set, but is largely ignored in relevant analyses. Here, I use Pacific halibut (*Hippoglossus stenolepis*) fishery as a case study to explore this effect, because it has both commercial and experimental (fishery-independent) data on hook spacing, and a survey-based CPUE series. It thus provides a unique opportunity to explore the effect of hook spacing and their impact on CPUE trends. Here, I explore this relationship using non-parametric and parametric relationships inside a spatially-explicit (geospatial) CPUE standardization model for the commercial data, and non-linear mixed-effects model for the experimental data. I found a clear non-linear spacing effect (i.e., hooks were less effective the closer they were), but accounting for space had a larger impact on CPUE trends than accounting for hook spacing. For this stock, it is likely the impact of hook spacing on CPUE was minimal due to the relatively constant average hook spacing over time. Regardless, historical and future trends in hook spacing can have important impacts on longline CPUE standardization, highlighting the value of collecting this information. Accounting for hook

spacing effects in other fisheries may improve the estimates of trends in relative abundance, and likely lead to better management.

3.2 Introduction

Catch per unit effort (CPUE) is a key source of information used to manage a wide range of commercially valuable species such as tunas, as well as vulnerable species like sharks [112]. CPUE is typically assumed to provide an index of population abundance N, that is robust for detecting trends and informing stock assessments provided that catchability q is constant through time and space, i.e., $CPUE = qN$. However, this assumption can fail for a variety of reasons. One important case is when catchability varies in time and space, such as when fish densities interact with fishermen behavior, and thus spatial patterns of catch [113, 114]. Another important case is when the unit of effort varies, such as with changing technological (e.g., gear) and economic factors or targeting strategies [115]. Either case undermines the comparability of CPUE between years and areas and can lead to effects like hyperdepletion or hyperstability [116], which complicates the interpretation of CPUE trends as accurately reflecting the true stock status trends (e.g., see [117, 118]). CPUE trends are thus typically standardized to remove effects other than changes in abundance, where possible, so they more accurately reflect changes in abundance [112, 115].

Standardizing CPUE from baited longline gear has the additional complexity that the probability of catching a fish, and thus catchability, depends on volitional (foraging) behavior which is affected by gear configuration and environmental variables [119]. This has been shown for important pelagic and demersal species caught by longline [120, 121, 122, 123]. Thus, it is important to consider the variation in configuration for longline gear in the CPUE standardization. Longline gear is a simple, but versatile, form of gear where baited hooks are attached to a mainline fixed at regular intervals ('fixed' gear), attached dynamically as it is deployed ('snap' gear), or attached at pre-determined points and deployed via an automated machine ('autoline' gear) [124]. Longline gear can be configured to target demersal species such as Pacific halibut (*Hippoglossus stenolepis*) and sablefish (*Anoplopoma fimbria*), as well

as pelagic species such as bigeye tuna (*Thunnus obesus*) [125]. Appropriately accounting for gear configuration in CPUE standardization is thus key for a wide range of important fisheries.

Although ostensibly simple, the interactions between longline gear and fish foraging behavior is complicated. A motivated fish must detect, locate, and then consume the bait, but each of these factors can strongly depend on varying environmental conditions such as temperature, turbidity, and light level, among others [119]. In addition, the density and size structure can affect fish behavior, such as when there is social facilitation with greater numbers of fish or a length hierarchy for feeding [120, 121]. These interactions complicate the definition of a unit of effort for longline gear, which would ostensibly be a hook (i.e., catch per hook). However, the spacing between hooks, as measured along the mainline, influences the foraging behavior of the target fish by affecting the region within which baits are detected, called the capture field or active space. Successfully accounting for the effect of hook spacing on effort could thus improve the CPUE standardization for longline gear.

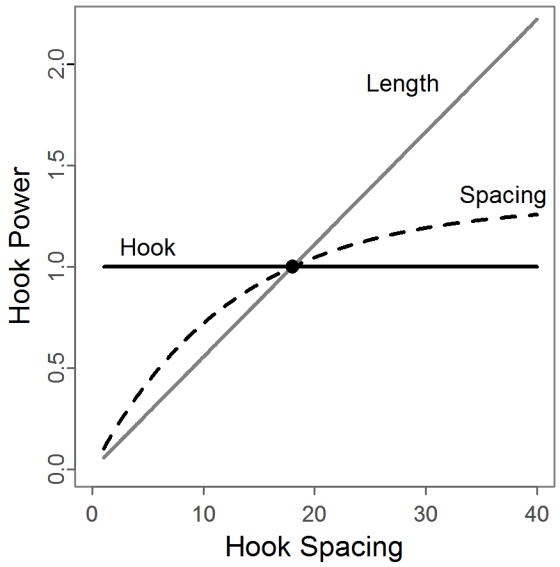

Figure 3.1: Stylized representations of three hypotheses for how the power of a hook changes with hook spacing, for a set with the same number of hooks but increasing total length and thus hook spacing. In the *hook* hypothesis, the hooks do not compete and thus the effective effort is the nominal hooks. In the *length* hypothesis, hooks compete at all spacings such that the length of the set is the effective effort. Lastly, the *spacing* hypothesis is intermediate and hooks compete at lower spacings only. Figure recreated from [126].

Three hypotheses have been proposed for how the capture field changes with hook spacing, which I refer to as 'spacing effect' (Fig. 3.1) [126]. Consider a hypothetical set with N hooks with varying hook spacings (and thus set length) fished at reasonable densities (e.g., hook saturation is not an issue) and uniformly distributed fish. In the *length* hypothesis, as spacing and set length decreases, the overlapping capture fields compete with each other, and catch per hook will decrease (e.g., [127]). In this case the length of the set would be the unit of effective effort. Alternatively, in the *hook* hypothesis, overlapping capture fields increase fish response, canceling out the effect of hook competition, and catch per hook is constant. In this case the unit of effort would be the number of hooks, and could occur when increased odor plumes from overlapping baits increased fish response from a wider area [128]. Lastly, the *spacing* hypothesis is intermediate, such that hooks spaced widely enough are effectively

independent, but hooks closer together would compete, to some degree, for the same fish. In this case, the nominal number of hooks are adjusted according to the hook spacing, to create effective hooks, and these are the unit of effort. Which of these hypotheses (hook spacing effect) is true is driven by the foraging ecology of the species of interest, not the gear itself, and is important for CPUE standardization because it defines what the effective unit of effort is for the gear.

The importance of correctly determining the effective effort in a longline data set depends on other properties of the gear. Consider the simplest case, where the same hook spacing is used consistently by all fishermen in all years. In this case, using the number of hooks or length of line will be equivalent up to a multiplicative factor which gets absorbed into catchability and leading to the same trends. However, ignoring effective effort when there is variation in hook spacing across either time, space, or fishermen, will bias the effort high or low for some sets and undermine the relationship between density and CPUE in unpredictable ways. Perhaps the most important case is when there is a temporal trend in the variation of hook spacing in gear deployment (e.g., using smaller and smaller spacings), which leads to a trend in the bias for effort (and thus CPUE), potentially creating a trend in apparent CPUE that is not related to abundance. This was the case in the Pacific halibut fishery with a notable shifts toward wider spacings from 1955 to 1970, resulting in misleading CPUE trends [129]. A similar concern remains in this fishery because of trends over time and space in the composition of the type of longline gear used, since each has a different spacing distribution (Fig. 3.2).

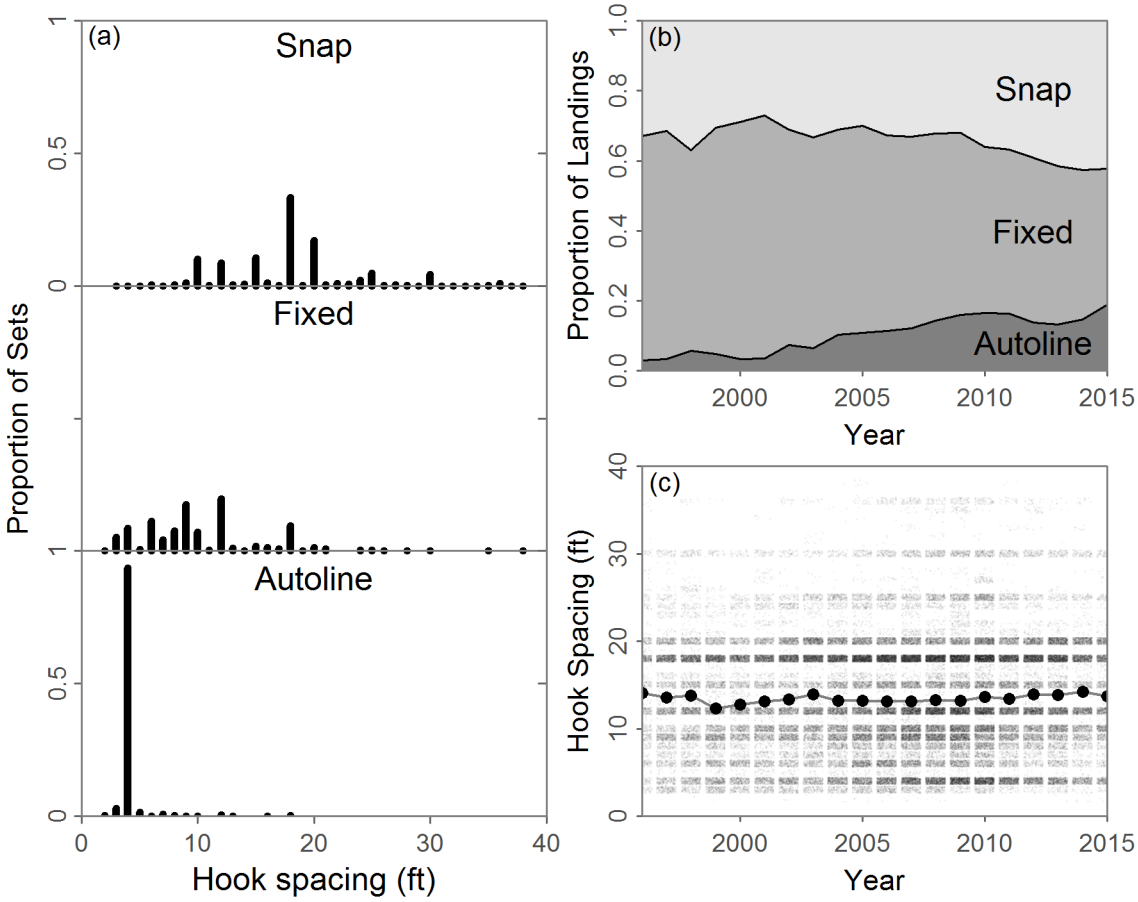

Figure 3.2: Properties of the fishery-dependent data (commercial catches). (a) The distribution of hook spacing within each of the three gear types. (b) Trends in proportion of catches by gear type by weight. (c) Annual distribution of hook spacing for all gear types (small points; jittered for clarity) and means (large points).

To investigate the spacing effect for Pacific halibut, Hamley and Skud (1978) initiated an experimental study (i.e., controlled fishing), but these data have insufficient samples at small spacings to adequately quantify this relationship over its current applied range. In contrast, recent commercial fishery data have wide variation in hook spacings, and provide an opportunity to quantify and contrast the spacing effect to that from the experimental data. In this study, I investigate the spacing effect for Pacific halibut, and its implication for standardized CPUE. First, I apply a spatially-explicit (spatiotemporal) standardization

model to commercial catch data to estimate standardized CPUE trends while simultaneously estimating the hook spacing effect. Then, I reanalyze the experimental data from Hamley and Skud (1978), and compare the two relationships and test whether the same information about the spacing effect is available in the commercial catch data. We conclude by discussing and demonstrating how these techniques can be used to improve CPUE standardization in longline fisheries.

3.3 Materials and Methods

3.3.1 Effective hooks

Here I hypothesize that as the distance to its neighbors varies, so will its power and thus effective effort of a hook (Fig. 3.1). I thus need a way to convert nominal hooks into effective hooks. The first step is to quantify the spacing effect with a function (f) that relates expected catch per hook with hook spacing. In this study, spacing data are reported as whole feet so I focus on discrete forms. I explore three possible relationships below.

Next, I adopted the approach taken by [126] and chose a reference spacing, and standardized relative to it to create what I could be thought of as relative hook power. I used 18 ft (5.5 m) as a reference in this study to maintain continuity with previous studies, and due to historical relevancy in the fishery. Relative hook power is unit-less and represents the relative ratio in efficiency between hooks fished at different spacings. For instance, a hook with a relative power of 0.5 at 10 ft indicates that I expect half the catch from that hook compared to if it were fished at 18 ft, all else being equal. I then calculated the number of effective hooks in a set as:

$$h(s) = h_{\text{nominal}} \cdot f(s)/f(18) \tag{3.1}$$

where h is the number of effective hooks, h_{nominal} the nominal (reported) number of hooks, s is the distance between hooks (in ft), and f is a mathematical function relating hook spacing and expected catch per hook (see below). Note that by definition sets fished with 18 ft

spacings will have equivalent nominal and effective hooks, but will be larger or smaller than the nominal hooks depending on the spacing.

To explore the shape of f I used three different forms. First, I used one without any effect of hook spacing:

$$f_{\text{constant}}(s) = 1 \tag{3.2}$$

This constant form ignores hook spacing and would be necessary, e.g., if spacings were not reported or available to the analyst.

I also fit a flexible random walk 'smoother' form, so that the shape of the hook spacing relationship could be elucidated with few *a priori* assumptions. I arbitrarily set the initial spacing effect at 1 ft, since it gets canceled out in calculating effective hooks in equation. Larger spacings were determined multiplying the previous spacing effect by a lognormal deviation. The deviations were modeled as random effects with a normal distribution: . Specifically, the smoother form is defined as:

$$f_{\text{smoother}}(s) = \begin{cases} 1 & \text{if } s = 1 \\ f(s-1)e^{\tau_d} & \text{otherwise} \end{cases} \tag{3.3}$$

Lastly, I fit a generalized version of the non-linear relationship used in Hamley and Skud (1978):

$$f_{\text{parametric}}(s) = \alpha \left[1 - \left(e^{-\beta_s s}\right)^{\lambda}\right] \tag{3.4}$$

As with the smoother form, the parameter α cancels out in the relative hook calculation and is thus fixed arbitrarily at $\alpha = 1$. We note that this formula represents the *spacing* hypothesis directly, but can also represent the *hook* and *length* hypotheses as special cases (as $\lambda \to \infty$ or $\beta \to \infty$, respectively). For this form, it is also possible to calculate an effective hook at infinite spacing, $h_\infty = h(\infty)$, analytically, which quantifies how close to independent hooks are at 18 ft.

Below I used the smoother form to explore the shape of the spacing effect, the parametric form to calculate relative abundance trends, and the constant form to test the effect of

ignoring hook spacing.

3.3.2 Analysis of fishery-dependent data

Data

These data come from International Pacific Halibut Commission (IPHC) commercial logbooks (summarized in [130]). The logbooks are required to be maintained, but only logbooks representing about 73% of landings were available in the IPHC database. The basic datum is a single longline 'set' fished by a commercial vessel. A set is a demersal longline consisting of sections (skates) of gear linked together. After soaking, the gear is retrieved and legal-sized halibut are retained and total weight for the set recorded. Data about the set, such as location, depth, gear properties, vessel, and date, are recorded in logbooks, and later collected and collated into a database at the IPHC. There are approximately 700,000 recorded sets over the time period 1991-2015, ranging from northern California to the Bering Sea. Obscured locations and summaries must be used here due to the confidential nature of these data.

These fishery-dependent data contain various information useful for CPUE standardization. The gear information includes the type of longline gear (fixed, snap, or autoline), hook spacing (or equivalently how many hooks were used), the length of gear deployed, and hook size. For most sets, the latitude and longitude at the beginning and end of the set were available. Although environmental factors are known to influence catch rates [119, 121], the only available environmental covariate is depth, which I also averaged over the beginning and end of each set.

For computational convenience, I narrowed the dataset down in several ways. First, I focused on data from the central Gulf of Alaska starting in 1996 because some previous geographic coordinates were recorded irregularly. I also filtered out sets without spatial coordinates or missing other key information. A small percentage of sets had zero catch (2.85%) and were excluded since they may represent targeting of other species, the reporting

rates were likely not consistent over time, and our focus was on hook spacing which requires non-zero catches. This step seemed reasonable because preliminary exploration of sets with zero catches had no apparent difference in hook spacing distributions. Since I used the midpoints of a set, including averaging depth, I also filtered out sets which were reported to be longer than 18 miles or had more than a 50 fa (91.44 m) difference in depth at the endpoints. Initial exploration suggested a relatively minor effect on the results due to the filtering. After filtering the data there were approximately 100,000 sets.

Spatiotemporal model

To explicitly account for space in my standardization, I fit spatiotemporal models [131, 132] to the commercial logbook data. These tools are used in many ecological fields, including estimating spatial densities of fish [133, 26]. In this modeling framework, the distribution of fish density is assumed to arise from unobserved environmental and biological factors. This density is assumed to vary smoothly in space and time and can be represented as a Gaussian random field, such that a finite set of points in space will have a multivariate normal distribution with spatial correlations captured by a covariance matrix.

Specifically, I model the relative fish density for set i, D_i, as:

$$D_i = \exp(\beta_0 + \beta_{y_i} y_i + \beta_{d_1} d_i + \beta_{d_2} d_i^2 + \omega_{c_i} + \epsilon_{c_i,y_i}) \tag{3.5}$$

for depth d and year y. The first four terms correspond to the typical component of a standardization, and the final two make up the spatial component. For the spatial component of the model, we adopted the spatial hierarchical statistical modelling approach (Cressie and Wikle 2015). That is, I fit a vector of random effects to all years (baseline spatial effects), ω, and separate vectors for each year (spatiotemporal effects), $\epsilon(y)$, with separate covariance matrices, but the same geostatistical decorrelation range κ. The index c_i dictates which spatial cell the density is in (see below). The result is a distinct spatial distribution of density for each year. Specifically, the random vectors were distributed as:

$$\omega \sim MVN(0, \Sigma_\omega) \tag{3.6}$$

$$\epsilon(y) \sim MVN(0, \Sigma_\epsilon) \tag{3.7}$$

Because the data are observed imperfectly, there is also an observation component of the model which accounts for expected catch, given density and external factors such as the number and spacing of hooks, gear type, and vessel. I define the expected catch for set i, μ_i, as:

$$\mu_i = h_i(s_i) \cdot D_i \cdot q_i \tag{3.8}$$

where D and f are defined as above, h_i the number of effective hooks, and $q_i = exp(g_i + \tau_{\nu_i})$ is the catchability for gear type g (i.e., fixed, snap or autoline), and vessel random effect τ_{ν_i} for vessel v_i, and where $\tau \sim N(0, \sigma_\tau)$. The vessel effect was included because I expect different vessels to have different fishing efficiencies and thus different expected catch, all else being equal. We further assume observed catches, C, have a log-normal distribution with estimated observation error σ_{obs}:

$$\log C_i \sim N(\log \mu_i, \sigma^2_{obs}). \tag{3.9}$$

Given the number of data points and resulting sizes of covariance matrices, this model is computationally infeasible. I therefore follow the lead of [26] and simplify the model in three ways. First, we approximate the random field by binning the data points into smaller regions or cells (defined by m "knots"), which reduces the dimensionality of the covariance matrices from n to approximately m, an approach known as predictive process modeling [134]. The placement of the knots was determined using the R function `kmeans` [45], which uses a clustering algorithm to partition the data such that the sum of squares from points to the assigned cluster centers is minimized. I then used the R package INLA [135] to create a mesh from the resulting cluster centers. The result is a distribution of cells within which all

data points are associated with the same spatial random effect. Initial exploration suggested that 2000 knots were sufficient to achieve convergence of the approximation (i.e., further increases resulted in no substantial changes to results).

Second, I reduce the number of parameters of the covariance matrices by using a Matérn semivariogram function with smoothness $\nu = 1$ [132]. The Matérn function relates the covariance between two points (or centers of cells) as a function of the distance between them, given range and variance parameters which are estimated from the data [136]. I further assumed isotropy and stationarity of the spatial process so that the orientation of the distance made no difference, and the spatiotemporal process, Σ_ϵ , was constant between years.

Lastly, I adopt the stochastic partial differential equation approach which converts the Gaussian random field into a Gaussian Markov random field [137, 135]. With this technique cells that are not directly neighbors are assumed to have zero covariance (i.e., be independent). By having non-zero covariance only for direct neighbors the inverse covariance matrix is sparse (has more off-diagonal zeros) which reduces computation [137]. These simplification techniques are widely used in geospatial modeling and greatly reduce the computational load while retaining the key properties of the spatial process of interest, making an analysis of 100,000 data points feasible.

Fully exploring a CPUE spatiotemporal standardization model for commercial catch data is beyond the scope of this paper. Here my focus is on accounting for enough of the biological and fishery properties to facilitate estimation of the hook spacing effect. I thus consider our spatiotemporal model a simplified model useful as a proof of concept, but note that there are independent estimates of relative abundance trends from a scientific survey over the same time and space [138], against which I compare and contrast my predictions.

Calculating CPUE trends

In contrast to other CPUE standardization models which explicitly model relative catch rates, my spatiotemporal model predicts densities in each cell. From these I multiply cell

density by its area (a_c), and then sum all cells to get annual total relative abundance, A_y:

$$A_y = \sum_c a_c D_{c,y}. \tag{3.10}$$

This calculation assumes that the process of selecting sampling sites is independent of the underlying biological process (density). This is true for surveys [26, 139], but here it is violated because captains are likely targeting areas with higher densities of fish. This is known as preferential sampling and can lead to biased inference [140]. Here my focus is on the hook spacing effect, and note that this is an open issue and analyses used for management should further investigate the bias and potentially mitigate it in the model [113, 141, 142, 139]. Here we calculate relative abundance trends with and without a spatial effect, and with and without a hook spacing effect to quantify the relative effects of these aspects.

3.3.3 Analysis of experimental data

Data

The 'experimental' data come from chartered commercial vessels fishing parallel sets of fixed gear with variable hook spacing (6-40 ft), repeated every day for 3-19 days at the same location [126]. These trips were repeated at different locations, in different years, but not always by different vessels. The catches varied by site, reflecting the underlying spatial variability in fish density (Fig. 3.3). As with the fishery-dependent data, I filtered out the sets with zero catches (2.7%), leaving 397 sets from 14 distinct locations.

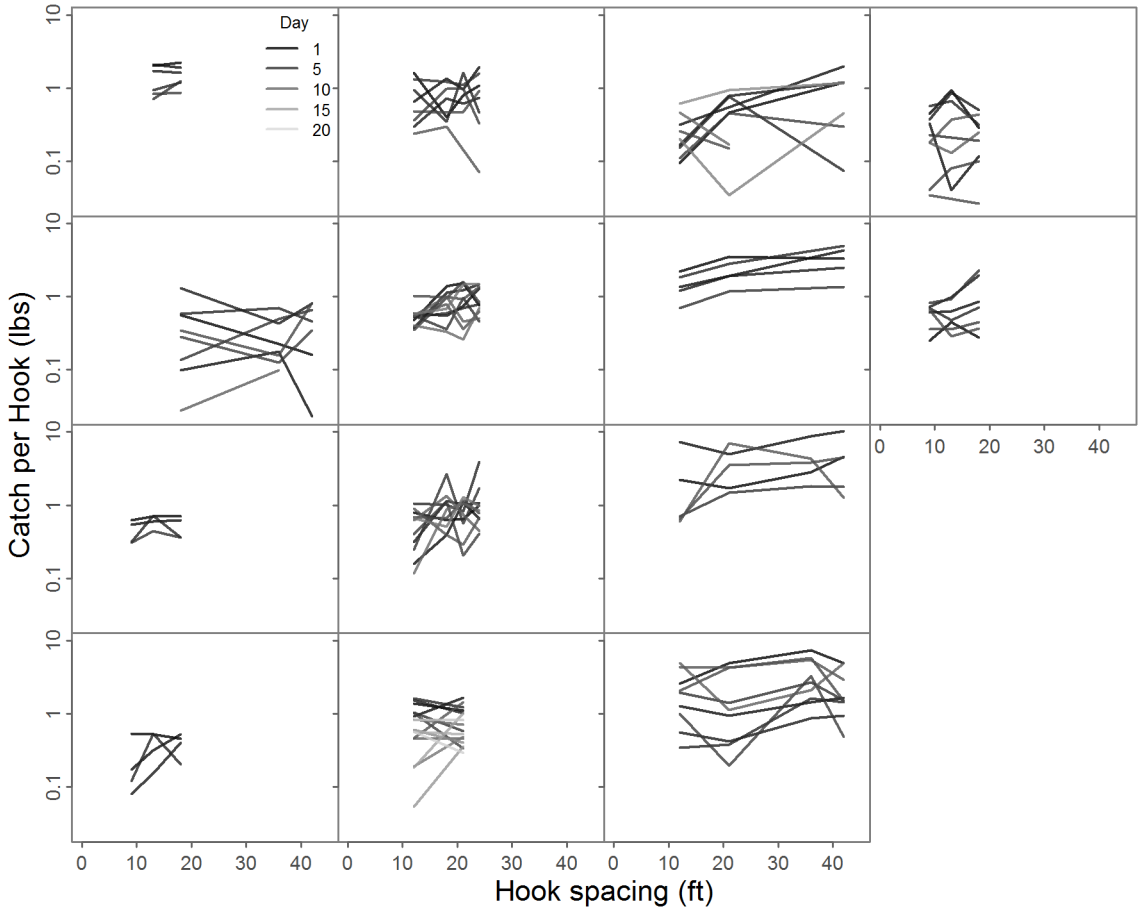

Figure 3.3: The raw data from Hamley and Skud (1978). Each panel is a separate site, and each line represents a series of sets fished at different spacings on the same day. Day number is colored. Sets with zero catch are removed.

These data differ from the commercial fishery-dependent data in that they were collected under a controlled sampling protocol. Nevertheless, the experimental data are unbalanced with respect to hook spacing, replicates, and vessels (Fig. 3.3). Local depletion is also a concern given that the same area was fished repeatedly, but for a variable number of days. Perhaps more importantly, there were few experiments with hook spacings at less than 10 ft, which is a commonly used spacing in the fishery nowadays.

Parametric hook spacing model

New methods and software now exist to take into account the complexities of the data which were largely ignored in the original least squares analysis. Specifically, I refit these data using non-linear mixed effects model that accounts for site-specific differences and local depletion. This model structure is widely used throughout ecology and fisheries, and better accounts for the data complexities and provides approximate uncertainty estimates about the fit and derived quantities [21].

As with the spatiotemporal model, eqn. (3.8), catch was predicted as a function of density, hook spacing, and catchability:

$$\mu_i = h_i(s_i) \cdot D_i \cdot q_i \tag{3.11}$$

In contrast to the spatiotemporal model, the densities of the sets are grouped by site, and I assume sites are distant enough to effectively be independent. Thus, I estimated site level densities as independent random effects, and included a local depletion term γ that such that density decreases exponentially with day d: $D_i = e^{\nu_{s_i} - \gamma d_i}$, where $\nu_s \sim N(\mu_\nu, \sigma_\nu^2)$ is the density for site s. No other information on environmental or gear quantities were available.

I only used the $f_{\text{parametric}}$ hook spacing form, and assumed that $\lambda = 1$ due to the lack of information at small spacings in the data. Since the data were collected in a controlled manner, I further set $q = 1$, such that the site level density effect captures catchability. Lastly, I assumed that observed catch is lognormally distributed, $\log C_i \sim N(\log \mu_i, \sigma_{s_i}^2)$, where σ_s is the site-specific observation random effect, assumed to be normally distributed: $\sigma_s \sim N(\theta, \sigma_\theta^2)$.

3.3.4 Model fitting

Both the spatiotemporal model and the parametric hook spacing model are non-linear hierarchical (mixed effects) models, containing both fixed and random effects. The most complex spatiotemporal model (using f_{smoother}) has 30 fixed effects and 59,254 total random effects

(1,116 for vessel effects, 2,765 for spatial, 55,300 for spatiotemporal, and 43 for smoother deviations). To fit these large, complex mixed effects models I used Template Model Builder (TMB; [143, 60]), which is a freely available tool that uses automatic differentiation to fit models using maximum marginal likelihood and random effect integration via the Laplace approximation [77]. Uncertainties in fixed effects were estimated using standard frequentist asymptotic assumptions, and derived quantities (such as hook power and relative abundance) via the Delta method, both of which are computed automatically by TMB.

INLA is a popular software tool for spatial models, and here I used it to generate inputs for the stochastic partial differential equation approach for my spatiotemporal model [135]. This model could have been fit with INLA [144], but by using TMB I had the convenience, and consistency, of using the same software platform for inference of all models.

3.4 Results

The spatiotemporal fishery-dependent analysis using the smoother hook spacing form showed a clear trend toward decreasing power of hooks with smaller spacings, albeit with much uncertainty at spacings wider than 30 ft (Fig. 3.4a). The parametric form estimated a maximum relative hook power, $h_\infty = 1.771$ (SE of 0.057). That is, a hypothetical set fished at spacings wide enough that hooks were independent would catch 1.771 times as much compared to at 18 ft. The parametric form for the experimental data was similar to that of the fishery-dependent ($h_\infty = 1.64$ (0.28); Table 2). These estimates suggest the hook spacing relationship asymptotes slower, and have lower hook power at smaller spacings than previously estimated (Fig. 3.4b). For the experimental model, the effect of local depletion (day of fishing) was positive and significant: $\gamma = 0.05$ (0.01). Overall, this model had much more uncertainty in the hook spacing effect, despite fixing $\lambda = 1$ (Table 3.1). In general, the parametric form from equation (3.4) matches the fits well, suggesting this form is reasonable for halibut.

Table 3.1: Model estimates and standard errors (parentheses) for the parametric model fit to the experimental data.

Parameter	Estimate (SE)
β_s	0.052 (0.015)
γ	0.048 (0.011)
λ	1.0 (-)
h_∞	1.636 (0.280)
θ	0.668 (0.062)
σ_θ	0.211 (0.050)
ϕ	-0.133 (0.199)
σ_ν	0.711 (0.142)

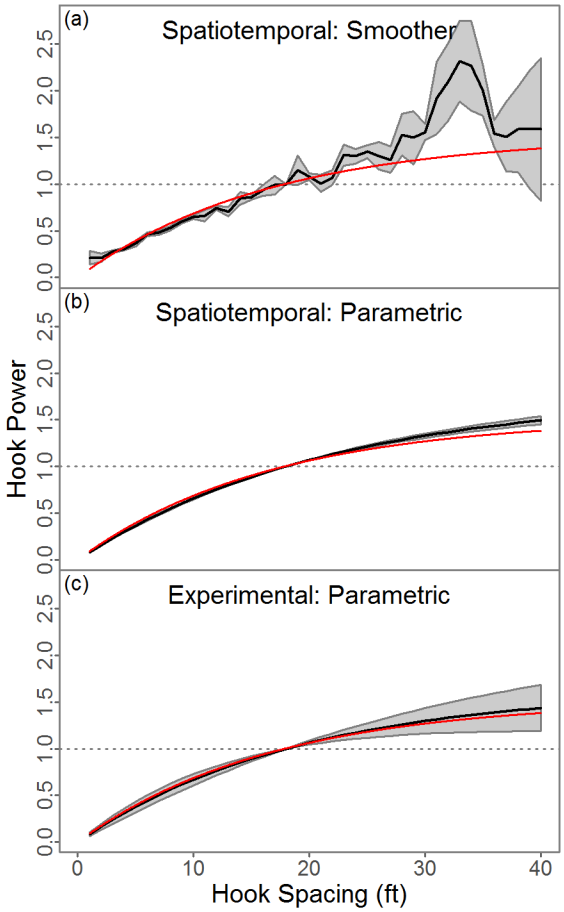

Figure 3.4: Estimated hook spacing effects for the smoother (a) and parametric forms from the spatiotemporal (b) and experimental model (c). Lines and shaded region show estimates and approximate 95% confidence interval, and red line shows historical parametric fit to experimental data from (Hamley and Skud 1978).

The spatiotemporal model estimates for the geostatistical properties were relatively insensitive to the form of hook spacing used. For instance, the variance of the spatiotemporal component (σ_ϵ) was 0.360, 0.342, and 0.345 (Table 3.2) for hook spacing effect of constant, smoother, and parametric forms, respectively. This pattern was not true for the observation error, σ_{obs} , where the models with the parametric form without space had a substantially larger estimate (0.772) than the spatiotemporal model (0.654). This 15% reduction in variance is expected as the spatiotemporal component explains variation in catch due to sets being proximate in time and space. The spatiotemporal residuals showed no clear spatial pattern (not shown – confidential), suggesting the model adequately captured those processes.

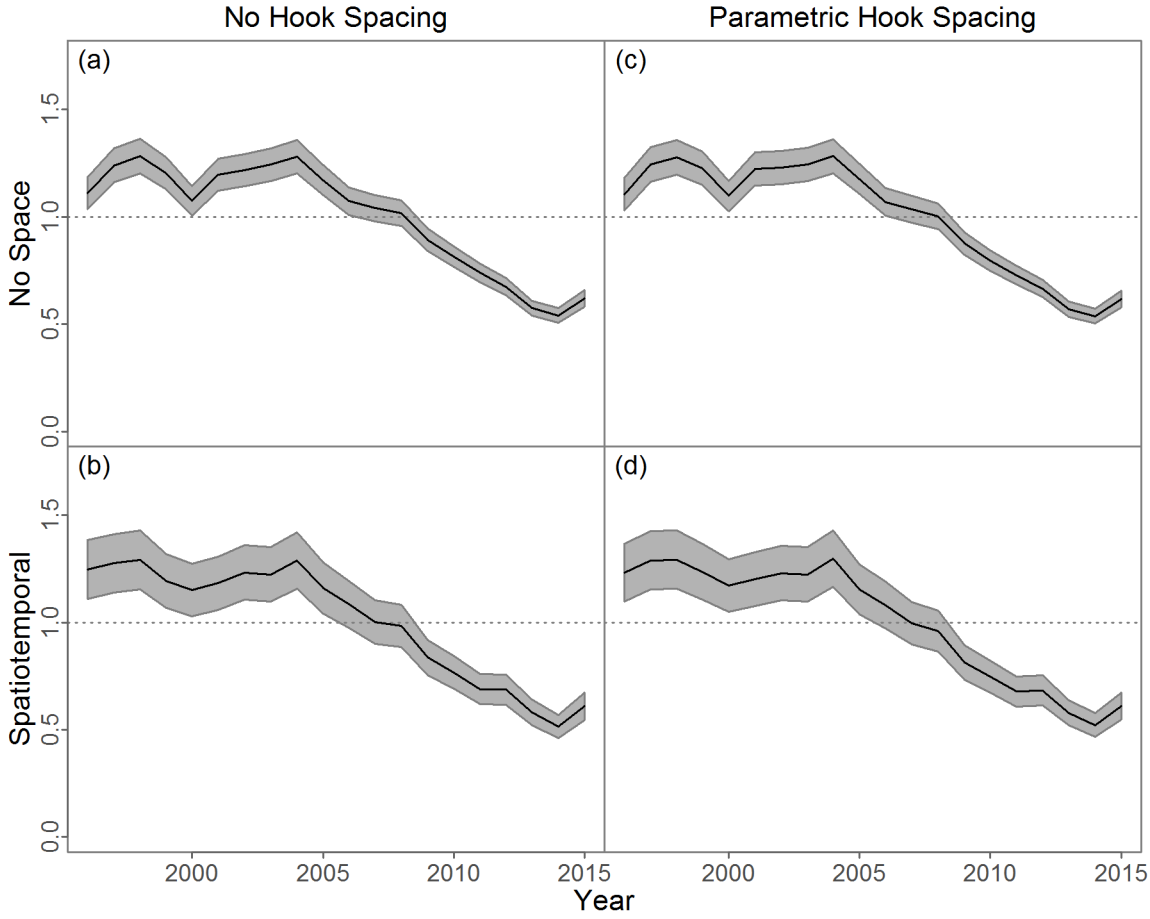

Figure 3.5: Effect of spatial component (rows) and hook spacing form (columns) on trends in relative abundance. Each panel is normalized by dividing by its mean. Lines and shaded region show estimates and approximate 95% confidence interval.

When using the fishery-dependent data to estimate trends in relative abundance the overall pattern was consistent, but there were some important differences (Fig. 3.5). All models predicted a relatively stable period from 1996 to 2004, a decline from 2005 to 2014, and a significant uptick in 2015. However, the uncertainty estimates for the spatiotemporal model were larger, particularly compared to the model without space or a hook spacing effect. There were some smaller annual differences when the hook spacing was not estimated, such as in 2007. However, in general the spatial effect had a much larger impact on predicted abundance trends than the effect of hook spacing. Compared to a trend estimated using

Table 3.2: Key model estimates and standard errors (parentheses) for models with and without space, and the parametric and smoother form for hook spacing. Depending on the model structure some parameters are not estimated, represented by (-), or the first level of a factor set to zero and thus there is no standard error. See appendix A for further results.

	Description	No Space Smoother	No Space Parametric	Spatiotemporal Smoother	Spatiotemporal Parametric
β_0	Global intercept	0.334 (0.037)	0.368 (0.035)	0.551 (0.060)	0.572 (0.058)
β_{d_1}	Linear effect of depth	3.66E-3 (1.89E-4)	3.65E-3 (1.89E-4)	8.54E-4 (2.72E-4)	1.03E-3 (2.74E-4)
β_{d_2}	Quadratic effect of depth	-7.46E-6 (8.23E-7)	-7.15E-6 (8.23E-7)	-3.99E-6 (9.95E-7)	-4.43E-6 (9.97E-7)
κ	Geostatistical range	-	-	0.399 (1.07E-2)	0.400 (1.07E-2)
σ_ϵ	Spatiotemporal variation	-	-	0.342 (5.68E-3)	0.345 (5.70E-3)
σ_ω	Spatial variation	-	-	0.370 (1.04E-2)	0.358 (1.00E-2)
g_1	Gear type: autoline	0 (-)	0 (-)	0 (-)	0 (-)
g_2	Gear type: fixed	0.341 (0.022)	0.348 (0.020)	0.249 (0.022)	0.269 (0.020)
g_3	Gear type: snap	0.093 (0.028)	0.080 (0.026)	0.083 (0.027)	0.099 (0.025)
σ_{obs}	Observation variance	0.770 (0.002)	0.772 (0.002)	0.653 (0.002)	0.654 (0.002)
σ_τ	Vessel variance	0.514 (0.013)	0.515 (0.013)	0.360 (0.010)	0.361 (0.010)
β_s	Parametric hook spacing	-	0.099 (0.515)	-	0.024 (0.040)
λ	Parametric hook spacing	-	0.567 (2.948)	-	1.925 (3.211)
σ_d	Smoother variation	0.142 (0.024)	-	0.140 (0.023)	-
h_∞	Power at infinite spacing	-	1.570 (0.042)	-	1.771 (0.057)

fishery-independent survey data (with constant hook spacing) over the same period of time and general area [145], my estimates had generally the same trend, but tended to have less annual changes and less uncertainty (Fig. 3.6).

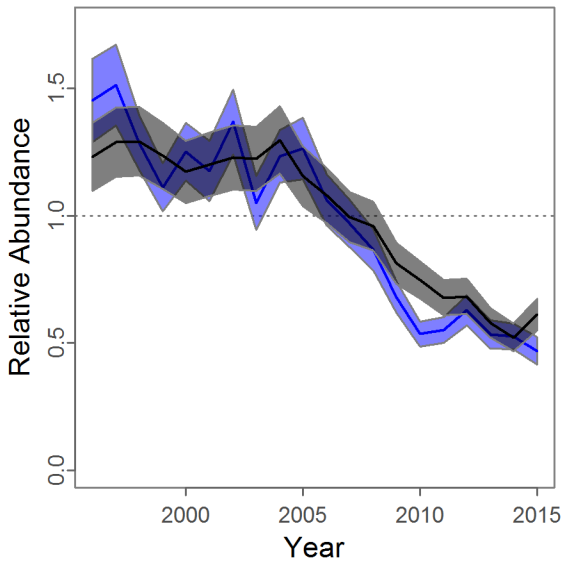

Figure 3.6: The relative abundance trend from the fishery-independent survey (blue) conducted under a controlled design (constant gear on a uniform grid, see [145]), compared to the parametric fit in the spatiotemporal model. Both are normalized to have mean of one. Lines and shaded region show estimates and approximate 95% confidence interval. 1996 and 1997 were unavailable for the survey series and thus left off for the spatiotemporal results.

3.5 Discussion

I found clear evidence for less effective hooks only at smaller spacings, supporting the hypothesis that nearby hooks compete for Pacific halibut. This implies that for CPUE analyses the relevant unit of effort is an effective hook. I also found that the parametric form, eqn. (3.4), was a reasonable approximation for this relationship. Further, the parametric fits to both the fishery-dependent and experimental data sets were fairly consistent, demonstrating this relationship can be estimated directly from commercial data, without the need for a controlled experiment. Estimating effective hooks in the CPUE standardization has the added

benefit that the uncertainty in the spacing effect is propagated into the trends of relative abundance. Lastly, despite a clear hook spacing effect, I found a limited impact on standardized CPUE trends examined here. This was likely due to the fact that although there has been a temporal shift to different gear types, on average the hook spacing has been relatively constant over the time period examined. Comparisons among other regulatory areas with systematic differences in gear usage may be much more important to the interpretation of Pacific halibut trends. Further, in other stocks managed with longline CPUE that do have significant temporal trends, ignoring hook spacing may mischaracterize abundance trends and lead to poor management decisions.

My results support the hypothesis that hooks compete with each other under, at least at the densities observed and conditioned on the specific foraging behavior of Pacific halibut. Since the data are collected exclusively *in situ*, these conclusions apply only at the population level, and are certainly affected by other factors. For example, I were not able to account for the effects of environment factors, size structure, or density on catch rates and thus caution against a biological interpretation of my results. I also caution against applying our estimates to other species or situations, as foraging behavior may vary widely and lead to fundamentally different relationships (Fig. 3.1). For instance, initial captures of sablefish do not affect subsequent captures leading to a random distribution of occupied hooks, while Pacific halibut tend to cluster [128]. Future lab experiments on Pacific halibut or other species, while controlling for environmental and other key factors, would provide valuable corroboration and further insights in the relationship between individual foraging behavior, hook competition, and the resulting population-level hook spacing effects.

The assessment of Pacific halibut uses CPUE that excludes snap and autoline gear due to concerns over confounding between gear type, hook spacing, and changes in density [76]. My analysis provides a method for including all gear types in future analyses and improving the information on which management is based. Although my analysis is specific to Pacific halibut, similar analyses for other stocks assessed, at least in part, with standardized longline CPUE could use a similar approach. For instance, hook spacing for sablefish is known to

be important from experiments, but is not consistently reported for commercial catches and thus cannot be directly used in the CPUE standardization [146]. Likewise, CPUE analyses for bigeye tuna account for hooks between floats and hooks per set, but the length of sets are unreported and thus the effect of hook spacing is unknown (e.g., [147]). My results demonstrate the potential value in collecting hook spacing for commercial longline catch data, and suggest incorporating this information in the future especially for stocks with temporal or spatial trends hook spacing over time.

Efforts to estimate fish stock status from longline CPUE trends while ignoring spatial effort have been widely criticized (e.g., see debates in [117, 113, 148]). As a consequence, incorporating spatial strata into standardizations is commonplace [112]. However, these improved methods still typically ignore spatial correlation among cells, and can be sensitive to cell resolution [149, 150]. One promising new method for accounting for space in standardizations is hierarchical spatiotemporal models [26]. These mixed effects models have become increasingly popular tools across a wide range of applications in fisheries science [22], and their application for spatiotemporal models provides a natural approach for dealing with the complexities of space in estimating fish densities. In contrast to data collected using a random design (e.g., surveys), the preferential sampling of commercial data (i.e., high density areas are targeted; see [140] remains an open issue when using these method. I did not attempt to address this issue in my simplified model, here used as a proof of concept and to investigate hook spacing effects, but note I were encouraged that my estimates closely matched a survey CPUE trend (Fig. 3.6). However, before using these methods for management, I suggest future studies more closely investigate the impacts of preferential sampling, in addition to other factors ignored here (e.g., zero catches and anisotropy), which may have an important influence on some stocks. I expect development of these models to continue being an active area of research, and will eventually be applied widely to analyze complex spatial fisheries data.

Trends in CPUE may not accurately reflect true trends in abundance due to a wide variety of confounding factors. Accounting for all such confounding factors is thus critical for

successful fisheries management, but is a difficult proposition and will be a source of continued research. For longline gear, in particular, the spacing between hooks clearly impacts the effective effort leading to observed catches. This highlights the value in collecting hook spacing data on longline sets, particularly if there is the potential for an annual trend in hook spacing as gear configuration evolves in a fishery. Fortunately, the effective effort implied by hook spacing can be estimated within a spatially-explicit CPUE standardization model fit to commercial catch data. Including this effective hook relationship will likely lead to improved trends in relative abundance, and hence better management for other species caught by longline.

3.6 Acknowledgements

I thank Jim Thorson for introducing us to spatiotemporal models, and Jim Thorson, Trevor Branch and Tim Essington for providing valuable feedback on an earlier version of this chapter. I also thank the commercial Pacific halibut fleet for their dedicated reporting of their log books, without which this analysis would not be possible. I received compensation for some of this work from the IPHC. This publication is partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA10OAR4320148 (2010-2015) and NA15OAR4320063 (2015-2020), Contribution No. 2017-075. This work was partially funded in by a grant from Washington Sea Grant, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award No. NA14OAR4170078.

CONCLUSION

Integrated stock assessments are complex, statistical non-linear mixed effects models that use multiple sources of data to assess alternative management actions on fish stocks. Increases in computational power and software developments have played a key role in the progression and improvement of these models in several ways. One example is when raw data need to be processed before being included in the assessment, such as a CPUE standardization. These analyses have traditionally been generalized linear models, which often do not adequately account for spatially-varying effort and fish densities. At the same time, the structural complexity of assessments has also increased, and it is now typical to use hundreds of parameters in the penalized maximum likelihood framework. However, despite these hardware and software improvements, these models remain difficult to fit in a Bayesian paradigm. In this dissertation, I used new software tools to improve the Pacific halibut index of abundance, and confronted long Bayesian run times for integrated assessments with a new algorithm and other custom-built software advances.

Trends in CPUE which accurately reflect true trends in abundance are often critical for successful fisheries management. Here, I fit a spatiotemporal CPUE standardization model with the new software package TMB. I found that explicitly accounting for space improved the estimated commercial longline CPUE trend, and these estimates were further improved by accounting for the spacing between hooks within the model. Collecting hook spacing data on commercial longline sets is most important when there is an annual trend in hook spacing, such as when gear configuration evolves in a fishery. Ignoring this effect can lead to inaccurate trends in CPUE which will negatively influence the ability of the assessment to accurately estimate trends in biomass. With TMB and hardware advances, it is now possible to advance CPUE standardizations beyond generalized linear models, and explicitly

incorporate spatiotemporal effects. This will likely improve estimates of a key data input to assessments, likely leading to better management of fish stocks.

Like CPUE standardization, Bayesian inference has also been limited by computational resources. A typical model development process is to built a model for maximum likelihood, adjusting the complexity based on expert knowledge, data availability, and the fishery's history. The final model is then taken "off the shelf," meaning explicit priors are not added, and a single MCMC chain is run, increasing the thinning rate until the model converges with sufficient effective samples. This stage can often take weeks or months to converge, which is prohibitive because models need to be sufficiently fast enough to allow sensitivity tests and other explorations required during the review process. Previously, these long run times were often attributed to the number of parameters, structural complexity of the model (e.g. time-varying parameters), or the antiquated Bayesian algorithm available in ADMB. Here, I confronted slow convergence by identifying problematic model properties, introducing a modern algorithm (the no-U-turn sampler), and streamlining the software workflow (including parallelization of MCMC chains). Surprisingly, I found the primary culprit for prohibitively long run times was poorly parameterized models, and not the model size or complexity, or Bayesian algorithm used. In this context, poor parameterization means trying to estimate parameters for which there is limited information in the data or priors to do so. The result is parameters stuck near bounds, fat tails, or posterior regions with high curvature (large gradients).

In the case studies I examined, many aspects of the models were an issue, but selectivity (the 'double-normal' curve in particular) was the most consistent in being over-parameterized. Simplifying the selectivity curves by adding priors or fixing parameters, a process known as regularization, made a substantial improvement when these issues occurred. A more subtle parameterization issue was observed with certain random effects which were strongly correlated, but in an anisotropic (non-linear) way. In this case, the default random-walk Metropolis algorithm in ADMB was unable to explore that part of the posterior, leading to biased parameter estimates. Alarmingly, this bias was not detectable by any traditional

MCMC diagnostics and would easily pass as being converged and appropriate to use for inference. Importantly, this means the default ADMB algorithm can produce biased inference which would not be solved by running longer chains. This is a serious issue that has not been addressed in the fisheries literature, let alone for stock assessments. Here, the bias only became apparent because the no-U-turn sampler better explored the posterior. This suggests that analysts should expect, and be cautious of, bias in posteriors for stock assessments when using the random-walk Metropolis algorithm. As such, I recommend the no-U-turn sampler as the default MCMC algorithm used for Bayesian inference for stock assessment models. This algorithm is available to all ADMB models compiled with my modified version of ADMB, which is freely available.

How, and if, to address issues like bias and poor parameterization are the topic of deep philosophical debate. Here, I emphasize that the entire process of stock assessment is subjective and driven by expert opinion: what data to use and how to weight them, which forms of selectivity to use, which key parameters must be fixed and at which values, etc. I argue that when performing statistical inference, whether frequentist or Bayesian, the model should be well defined and meet its underlying statistical assumptions. Thus, solutions like re-parameterizing, adding realistic priors, or simplifying the model (e.g., changing to a selectivity form with fewer parameters) based on expert opinion are justified and likely necessary for most stock assessments. I call this process regularization and posit that it is a key step in taking an assessment "off the shelf" and getting Bayesian inference in a reasonable time.

Although not the focus here, specifying priors is a key part of Bayesian analysis. Future work estimating informative priors from meta-analyses or assessments of similar stocks will continue to be a valuable source of information to contribute to inference, and will also likely help reduce run times. In particular, studies examining reasonable priors for complex selectivity forms like the double-normal in Stock Synthesis are needed, as the current defaults are not designed for Bayesian inference, but rather to stabilize optimization in maximum likelihood analyses. Alternatively, new selectivity forms could be developed to have parameters such that priors are easier to specify.

Despite regularization, many computational issues may remain. For instance, there is no clear solution to the difficult correlation between certain random effects. Although the no-U-turn sampler explored this space better, it still challenged the algorithm and increased run times substantially. The core issue is that the pairwise correlation occurs in sequential parameters, creating a difficult shape for an algorithm to explore in higher dimensions. Extensions of NUTS, particularly Riemannian Manifold HMC, are designed precisely to deal with such difficult geometries and thus are a promising extension of the work here. Further, there run time per evaluation among models can vary widely, and this directly affects run time for Bayesian (and frequentist) inference. Thus, constructing models to more efficiently project population dynamics and calculate log-likelihoods or log-densities is another path to decreasing run time for Bayesian inference.

In this dissertation, I demonstrated that hardware and software advances can improve inference in fisheries stock assessment in two ways. First, more sophisticated spatiotemporal models are now possible which allow for improved processing of key data inputs. Second, run times for integrated stock assessments can be reduced substantially through the use of a new Bayesian algorithm and taking advantage of multiple cores with parallelization. Specifically, run time for off-the-shelf models can be reduced by several orders of magnitude with the following steps: (1) Regularization of models (one to two orders), (2) Applying ten parallel chains, which was straightforward on a desktop computer for all of the models examined (one order of magnitude), (3) Using the NUTS algorithm (one order of magnitude, depending on the geometry of the posterior). Thus, through the advice and software provided in this dissertation, overnight analyses are likely feasible (though not trivial) for even the most challenging models. This advancement of Bayesian methods for stock assessments provides a tool for analysts to better understand the consequences of management decisions and will improve the management of fisheries.

BIBLIOGRAPHY

[1] R. Hilborn and C. J. Walters. *Quantitative Fisheries Stock Assessment: choice, Dynamics and Uncertainty*. Springer Science & Business Media, 1992.

[2] M. B. Schaefer. Some considerations of population dynamics and economics in relation to the management of the commercial marine fisheries. *Journal of the Fisheries Board of Canada*, 14:669–681, 1957.

[3] W. Doubleday. A least squares approach to analyzing catch at age data. *Int. Comm. Northwest Atl. Fish. Res. Bull*, 12:69–81, 1976.

[4] D. Fournier and C. P. Archibald. A general theory for analyzing catch at age data. *Canadian Journal of Fisheries and Aquatic Sciences*, 39:1195–1207, 1982.

[5] M. N. Maunder and A. E. Punt. A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, 142:61–74, 2013.

[6] A. Griewank. On automatic differentiation. *Mathematical Programming: Recent Developments and Applications*, 6:83–107, 1989.

[7] D. A. Fournier, H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen, and J. Sibert. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods & Software*, 27:233–249, 2012.

[8] R. D. Methot and C. R. Wetzel. Stock Synthesis: a biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research*, 142:86–99, 2013.

[9] R. Hilborn, M. Maunder, A. Parma, B. Ernst, J. Payne, and P. Starr. Coleraine: a generalized age structured stock assessment model. Version 2.0, Rep. SAFS-UW-0116, University of Washington, Seattle, 2003.

[10] D. A. Fournier, J. Hampton, and J. R. Sibert. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. *Canadian Journal of Fisheries and Aquatic Sciences*, 55:2105–2116, 1998.

[11] B. Bull, R. Francis, A. Dunn, A. McKenzie, D. Gilbert, M. Smith, R. Bian, and D. Fu. CASAL (C++ algorithmic stock assessment laboratory): CASAL user manual v2, 2005.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[13] S. M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, pages 359–363, 1986.

[14] M. K. McAllister, E. K. Pikitch, A. E. Punt, and R. Hilborn. A Bayesian approach to stock assessment and harvest decisions using the Sampling/Importance Resampling Algorithm. *Canadian Journal of Fisheries and Aquatic Sciences*, 51:2673–2687, 1994.

[15] R. Meyer and R. B. Millar. Bayesian stock assessment using a state-space implementation of the delay difference model. *Canadian Journal of Fisheries and Aquatic Sciences*, 56:37–52, 1999.

[16] R. B. Millar and R. Meyer. Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings within-Gibbs sampling. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 49:327–342, 2000.

[17] R. Hilborn, E. K. Pikitch, and M. K. McAllister. A Bayesian estimation and decision analysis for an age-structured model using biomass survey data. *Fisheries Research*, 19:17–30, 1994.

[18] M. Liermann and R. Hilborn. Depensation in fish stocks: a hierarchic Bayesian meta-analysis. *Canadian Journal of Fisheries and Aquatic Sciences*, 54:1976–1984, 1997.

[19] A. M. Ellison. Bayesian inference in ecology. *Ecology Letters*, 7:509–520, 2004.

[20] M. K. McAllister, E. K. Pikitch, and E. A. Babcock. Using demographic methods to construct Bayesian priors for the intrinsic rate of increase in the Schaefer model and implications for stock rebuilding. *Canadian Journal of Fisheries and Aquatic Sciences*, 58:1871–1890, 2001.

[21] J. A. Royle and R. M. Dorazio. *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, 2008.

[22] J. T. Thorson and C. Minto. Mixed effects: a unifying framework for statistical modelling in fisheries biology. *ICES Journal of Marine Science*, 72:1245–1256, 2014.

[23] R. Hilborn and M. Liermann. Standing on the shoulders of giants: learning from experience in fisheries. *Reviews in Fish Biology and Fisheries*, 8:273–283, 1998.

[24] K. Sainsbury. Effect of individual variability on the von Bertalanffy growth equation. *Canadian Journal of Fisheries and Aquatic Sciences*, 37:241–247, 1980.

[25] R. Francis. Maximum likelihood estimation of growth and growth variability from tagging data. *New Zealand Journal of Marine and Freshwater Research*, 22:43–51, 1988.

[26] J. T. Thorson, A. O. Shelton, E. J. Ward, and H. J. Skaug. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. *ICES Journal of Marine Science*, 72:1297–1310, 2015.

[27] A. E. Punt and R. Hilborn. Fisheries stock assessment and decision analysis: the Bayesian approach. *Reviews in Fish Biology and Fisheries*, 7:35–63, 1997.

[28] M. K. McAllister and G. P. Kirkwood. Bayesian stock assessment: a review and example application using the logistic model. *ICES Journal of Marine Science*, 55:1031–1060, 1998.

[29] M. D. Hoffman and A. Gelman. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

[30] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC Press, 2011.

[31] M. B. Hooten and N. T. Hobbs. A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85:3–28, 2015.

[32] B. M. Bolker, B. Gardner, M. Maunder, C. W. Berg, M. Brooks, L. Comita, E. Crone, S. Cubaynes, T. Davies, P. de Valpine, J. Ford, O. Gimenez, M. Kery, E. J. Kim, C. Lennert-Cody, A. Magnusson, S. Martell, J. Nash, A. Nielsen, J. Regetz, H. Skaug, and E. Zipkin. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. *Methods in Ecology and Evolution*, 4:501–512, 2013.

[33] N. Cressie, C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19:553–570, 2009.

[34] R.M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

[35] A. Gelman, D. Lee, and J. Q. Guo. Stan: a probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40:530–543, 2015.

[36] Stan Development Team. Stan modeling language users guide and reference manual, version 2.11.0, 2016.

[37] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, A. Riddell, J. Q. Guo, P. Li, and A. Riddell. Stan: a probabilistic programming language. *Journal of Statistical Software*, 76:1–29, 2017.

[38] M. Plummer. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 124:125, 2003.

[39] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[40] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

[41] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222, 1987.

[42] M. Betancourt. Identifying the optimal integration time in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.00225*, 2016.

[43] M. Betancourt, S. Byrne, and M. Girolami. Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1411.6669*, 2014.

[44] M. Betancourt, S. Byrne, S. Livingstone, and M. Girolami. The geometric foundations of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1410.5110*, 2014.

[45] R Core Team. R: a language and environment for statistical computing, 2016.

[46] O. Papaspiliopoulos, G. O. Roberts, and M. Skold. A general framework for the parametrization of hierarchical models. *Statistical Science*, 22:59–73, 2007.

[47] M. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications*, 79:30, 2015.

[48] J. Schnute. A versatile growth model with statistically stable parameters. *Canadian Journal of Fisheries and Aquatic Sciences*, 38:1128–1140, 1981.

[49] M. Kry and M. Schaub. *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press, 2012.

[50] F. Korner-Nievergelt, T. Roth, S. von Felten, J. Gulat, B. Almasi, and P. Korner-Nievergelt. *Bayesian data analysis in ecology using linear models with R, BUGS, and Stan: including comparisons to frequentist statistics*. Academic Press, 2015.

[51] R. L. Grant, D. C. Furr, B. Carpenter, and A. Gelman. Fitting Bayesian item response models in Stata and Stan. *arXiv preprint arXiv:1601.03443*, 2016.

[52] M. Betancourt. Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1604.00695*, 2016.

[53] S. Livingstone, M. Betancourt, S. Byrne, and M. Girolami. On the geometric ergodicity of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.08057*, 2016.

[54] D. Dail and L. Madsen. Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics*, 67:577–587, 2011.

[55] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 73:123–214, 2011.

[56] M. Betancourt. Generalizing the no-U-turn sampler to Riemannian manifolds. *arXiv preprint arXiv:1304.1920*, 2013.

[57] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*, 2016.

[58] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–413, 2017.

[59] J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5:65–80, 2010.

[60] K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. Bell. TMB: automatic differentiation and Laplace approximation. *arXiv preprint arXiv:1509.00660*, 2015.

[61] I. J. Stewart, A. C. Hicks, I. G. Taylor, J. T. Thorson, C. Wetzel, and S. Kupschus. A comparison of stock assessment uncertainty estimates using maximum likelihood and Bayesian methods implemented with the same model framework. *Fisheries Research*, 142:37–46, 2013.

[62] M. Schaub and F. Abadi. Integrated population models: a novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology*, 152:227–237, 2011.

[63] C. S. Szuwalski and J. Turnock. A stock assessment for eastern Bering Sea snow crab. Available at `https://www.npfmc.org/wp-content/PDFdocuments/resources/SAFE/CrabSAFE/2016CrabSAFE_final.pdf`, 2016.

[64] M. L. Muradian, T. A. Branch, S. D. Moffitt, and P.-J. F. Hulson. Bayesian stock assessment of Pacific herring in Prince William Sound, Alaska. *Plos One*, 12:e0172153, 2017.

[65] A. E. Punt, T. Huang, and M. N. Maunder. Review of integrated size-structured models for stock assessment of hard-to-age crustacean and mollusc species. *ICES Journal of Marine Science: Journal du Conseil*, 70:16–33, 2013.

[66] R. B. Millar and R. J. Fryer. Estimating the size-selection curves of towed gears, traps, nets and hooks. *Reviews in Fish Biology and Fisheries*, 9:89–116, 1999.

[67] D. B. Sampson and R. D. Scott. A spatial model for fishery age-selection at the population level. *Canadian Journal of Fisheries and Aquatic Sciences*, 68:1077–1086, 2011.

[68] A. E. Punt, F. Hurtado-Ferro, and A. R. Whitten. Model selection for selectivity in fisheries stock assessments. *Fisheries Research*, 158:124–134, 2014.

[69] G. Aarts and J. Poos. Comprehensive discard reconstruction and abundance estimation using flexible selectivity functions. *ICES Journal of Marine Science: Journal du Conseil*, 66:763–771, 2009.

[70] J. T. Thorson and I. G. Taylor. A comparison of parametric, semi-parametric, and non-parametric approaches to selectivity in age-structured assessment models. *Fisheries Research*, 158:74–83, 2014.

[71] D. Butterworth, J. Ianelli, and R. Hilborn. A statistical model for stock assessment of southern bluefin tuna with temporal changes in selectivity. *African Journal of Marine Science*, 25:331–361, 2003.

[72] J. T. Thorson, C. C. Monnahan, and J. M. Cope. The potential impact of time-variation in vital rates on fisheries management targets for marine fishes. *Fisheries Research*, 169:8–17, 2015.

[73] K. F. Johnson, C. C. Monnahan, C. R. McGilliard, K. A. Vert-pre, S. C. Anderson, C. J. Cunningham, F. Hurtado-Ferro, R. R. Licandeo, M. L. Muradian, K. Ono, C. S.

Szuwalski, J. L. Valero, A. R. Whitten, and A. E. Punt. Time-varying natural mortality in fisheries stock assessment models: identifying a default approach. *ICES Journal of Marine Science*, 72:137–150, 2014.

[74] J. T. Thorson, A. C. Hicks, and R. D. Methot. Random effect estimation of time-varying factors in Stock Synthesis. *ICES Journal of Marine Science: Journal du Conseil*, 72:178–185, 2015.

[75] J. T. Thorson and C. Wetzel. The status of canary rockfish (*Sebastes pinniger*) in the California Current. Availble at `http://www.cio.noaa.gov/services_programs/prplans/pdfs/ID308_FinalProduct_CanaryRockfish_2016.pdf`, 2015.

[76] I. J. Stewart, C. C. Monnahan, and S. Martell. Assessment of the Pacific halibut stock at the end of 2015. Available at `http://iphc.int/publications/rara/2015/RARA2015_12Assessment.pdf`, 2016.

[77] H. J. Skaug and D. A. Fournier. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, 51:699–709, 2006.

[78] R. Hilborn. The state of the art in stock assessment: where we are and where we are going. *Scientia Marina*, 67:15–20, 2003.

[79] A. Magnusson, A. E. Punt, and R. Hilborn. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. *Fish and Fisheries*, 14:325–342, 2013.

[80] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[81] R. Hilborn and M. Mangel. *The ecological detective: confronting models with data*, volume 28. Princeton University Press, 1997.

[82] S. P. Wang, M. N. Maunder, K. R. Piner, A. Aires-da Silva, and H. H. Lee. Evaluation of virgin recruitment profiling as a diagnostic for selectivity curve structure in integrated stock assessment models. *Fisheries Research*, 158:158–164, 2014.

[83] P. de Valpine. Shared challenges and common ground for Bayesian and classical analysis of hierarchical statistical models. *Ecological Applications*, 19:584–588, 2009.

[84] C. C. Monnahan, M. L. Muradian, and P. T. Kuriyama. A guide for Bayesian analysis in AD Model Builder. Available at `http://www.admb-project.org/developers/mcmc/mcmc-guide-for-admb.pdf`, 2014.

[85] M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: convergence diagnostics and out analysis for MCMC. *R News*, 6:7–11, 2006.

[86] J. T. Thorson and J. M. Cope. Uniform, uninformed or misinformed?: The lingering challenge of minimally informative priors in data-limited Bayesian stock assessments. *Fisheries Research*, 194:164–172, 2017.

[87] S. Van Dongen. Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Bioliology*, 242:90–100, 2006.

[88] S. R. Lele and B. Dennis. Bayesian methods for hierarchical models: are ecologists making a Faustian bargain? *Ecological Applications*, 19:581–584, 2009.

[89] A. J. R. Cotter, L. Burt, C. G. M. Paxton, C. Fernandez, S. T. Buckland, and J. X. Pax. Are stock assessment methods too complicated? *Fish and Fisheries*, 5:235–254, 2004.

[90] T. A. Branch, K. Matsuoka, and T. Miyashita. Evidence for increases in Antarctic blue whales based on Bayesian modelling. *Marine Mammal Science*, 20:726–754, 2004.

[91] M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[92] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[93] C. C. Monnahan, J. T. Thorson, and T. A. Branch. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 10.1111/2041-210x.12681, 2016.

[94] C. C. Monnahan, K. Ono, S. C. Anderson, M. B. Rudd, A. C. Hicks, F. Hurtado-Ferro, K. F. Johnson, P. T. Kuriyama, R. R. Licandeo, C. C. Stawitz, I. G. Taylor, and J. L. Valero. The effect of length bin width on growth estimation in integrated age-structured stock assessments. *Fisheries Research*, 180:103–112, 2016.

[95] R. D. Methot. User manual for Stock Synthesis. Version 3.24s. Available at `http://www.st.nmfs.noaa.gov/Assets/science_program/SS_User_Manual_3.24s.pdf`, 2015.

[96] James T. Thorson and Jason M. Cope. Uniform, uninformed or misinformed?: The lingering challenge of minimally informative priors in data-limited Bayesian stock assessments. *Fisheries Research*, 194:164 – 172, 2017.

[97] Stan Development Team. RStan: the R interface to Stan. R package version 2.14.1. `http://mc-stan.org`, 2016.

[98] J. Gabry. bayesplot: plotting for Bayesian models. R package version 1.2.0, `http://mc-stan.org/`, 2017.

[99] Stan Development Team. shinystan: interactive visual and numerical diagnostics and posterior analysis for Bayesian models, 2017.

[100] C. C. Monnahan. adnuts: no-U-turn sampling for ADMB and TMB models. Available at `www.github.com/colemonnahan/adnuts`, 2017.

[101] S. C. Anderson, C. C. Monnahan, K. F. Johnson, K. Ono, and J. L. Valero. ss3sim: an R package for fisheries stock assessment simulation with Stock Synthesis. *PLOS ONE*, 9:e92725, 2014.

[102] C. J. Grandin, A. C. Hicks, A. M. Berger, N. Edwards, I. G. Taylor, and S. Cox. Status of the Pacific hake (whiting) stock in U.S. and Canadian waters in 2016, 2016.

[103] C. A. Akselrud, A. E. Punt, and L. Cronin-Fine. Exploring model structure uncertainty using a general stock assessment framework: the case of Pacific cod in the Eastern Bering Sea. *Fisheries Research*, 193:104–120, 2017.

[104] A. E. Punt, C. A. Akselrud, and L. Cronin-Fine. The effects of applying mis-specified age- and size-structured models. *Fisheries Research*, 188:58–73, 2017.

[105] I. J. Stewart and C. C. Monnahan. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. *Fisheries Research*, 10.1016/j.fishres.2016.06.018, 2016.

[106] X. He, J. C. Field, D. E. Pearson, and L. S. Lefebvre. Age sample sizes and their effects on growth estimation and stock assessment outputs: three case studies from U.S. West Coast fisheries. *Fisheries Research*, 180:92–102, 2016.

[107] P. T. Kuriyama, K. Ono, F. Hurtado-Ferro, A. C. Hicks, I. G. Taylor, R. R. Licandeo, K. F. Johnson, S. C. Anderson, C. C. Monnahan, M. B. Rudd, C. C. Stawitz, and J. L. Valero. An empirical weight-at-age approach reduces estimation bias compared to modeling parametric growth in integrated, statistical stock assessment models when growth is time varying. *Fisheries Research*, 180:119–127, 2016.

[108] I. J. Stewart and S. J. D. Martell. Reconciling stock assessment paradigms to better inform fisheries management. *ICES Journal of Marine Science: Journal du Conseil*, 10.1093/icesjms/fsv061, 2015.

[109] A. Nielsen and C. W. Berg. Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research*, 158:96–101, 2014.

[110] C. W. Berg and A. Nielsen. Accounting for correlated observations in an age-based state-space stock assessment model. *ICES Journal of Marine Science*, 73:1788–1797, 2016.

[111] S. Ralston, A. E. Punt, O. S. Hamel, J. D. DeVore, and R. J. Conser. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. *Fishery Bulletin*, 109:217–232, 2011.

[112] M. N. Maunder and A. E. Punt. Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, 70:141–159, 2004.

[113] C. Walters. Folly and fantasy in the analysis of spatial catch rate data. *Canadian Journal of Fisheries and Aquatic Sciences*, 60:1433–1436, 2003.

[114] T. A. Branch, R. Hilborn, A. C. Haynie, G. Fay, L. Flynn, J. Griffiths, K. N. Marshall, J. K. Randall, J. M. Scheuerell, E. J. Ward, and M. Young. Fleet dynamics and fishermen behavior: lessons for fisheries managers. *Canadian Journal of Fisheries and Aquatic Sciences*, 63:1647–1668, 2006.

[115] J. Bishop. Standardizing fishery-dependent catch and effort data in complex fisheries with technology change. *Reviews in Fish Biology and Fisheries*, 16:21–38, 2006.

[116] S. J. Harley, R. A. Myers, and A. Dunn. Is catch-per-unit-effort proportional to abundance? *Canadian Journal of Fisheries and Aquatic Sciences*, 58:1760–1772, 2001.

[117] R. A. Myers and B. Worm. Rapid worldwide depletion of predatory fish communities. *Nature*, 423:280–283, 2003.

[118] T. Polacheck. Tuna longline catch rates in the Indian Ocean: did industrial fishing

result in a 90% rapid decline in the abundance of large predatory species? *Marine Policy*, 30:470–482, 2006.

[119] A. W. Stoner. Effects of environmental variables on fish feeding ecology: implications for the performance of baited fishing gear and stock assessment. *Journal of Fish Biology*, 65:1445–1471, 2004.

[120] A. Stoner and M. Ottmar. Fish density and size alter Pacific halibut feeding: implications for stock assessment. *Journal of Fish Biology*, 64:1712–1724, 2004.

[121] A. W. Stoner, M. L. Ottmar, and T. P. Hurst. Temperature affects activity and feeding motivation in Pacific halibut: implications for bait-dependent fishing. *Fisheries Research*, 81:202–209, 2006.

[122] K. A. Bigelow and M. N. Maunder. Does habitat or depth influence catch rates of pelagic species? *Canadian Journal of Fisheries and Aquatic Sciences*, 64:1581–1594, 2007.

[123] P. Ward. Empirical estimates of historical variations in the catchability and fishing power of pelagic longline fishing gear. *Reviews in Fish Biology and Fisheries*, 18:409–426, 2008.

[124] A. Bjordal and S. Løkkeborg. *Longlining*. Fishing News Books, 1996.

[125] S. Løkkeborg, A. Fernö, and O. B. Humborstad. Fish behavior in relation to longlines. *Behavior of Marine Fishes: Capture Processes and Conservation Challenges*, pages 105–141, 2010.

[126] J. M. Hamley and B. E. Skud. Factors affecting longline catch and effort: II. Hook-Spacing. Available at `http://www.iphc.int/publications/scirep/SciReport0064.pdf`, 1978.

[127] D. M. Eggers, N. A. Rickard, D. G. Chapman, and R. R. Whitney. A methodology for estimating area fished for baited hooks and traps along a ground line. *Canadian Journal of Fisheries and Aquatic Sciences*, 39:448–453, 1982.

[128] M. F. Sigler. Abundance estimation and capture of sablefish (*Anoplopoma fimbria*) by longline gear. *Canadian Journal of Fisheries and Aquatic Sciences*, 57:1270–1283, 2000.

[129] B. E. Skud. A reassessment of effort in the halibut fishery. Available at `http://www.iphc.int/publications/scirep/SciReport0054.pdf`, 1972.

[130] C. C. Monnahan and I. J. Stewart. Evaluation of commercial logbook records: 1991-2013. Available at `http://www.iphc.int/publications/rara/2014/rara2014_14commlog_revision.pdf`, 2014.

[131] J. S. Clark. *Models for ecological data: an introduction*, volume 11. Princeton University Press, 2007.

[132] N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

[133] A. O. Shelton, J. T. Thorson, E. J. Ward, and B. E. Feist. Spatial semiparametric models improve estimates of species abundance and distribution. *Canadian Journal of Fisheries and Aquatic Sciences*, 71:1655–1666, 2014.

[134] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:825–848, 2008.

[135] F. Lindgren and H. Rue. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63:1–25, 2015.

[136] J. A. Royle and C. K. Wikle. Efficient statistical mapping of avian count data. *Environmental and Ecological Statistics*, 12:225–243, 2005.

[137] F. Lindgren, H. Rue, and J. Lindstrom. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 73:423–498, 2011.

[138] I. J. Stewart and C. C. Monnahan. Overview of data sources for the Pacific halibut stock assessment and related analyses. Available at `http://iphc.int/publications/rara/2015/RARA2015_12Assessment.pdf`, 2016.

[139] J. T. Thorson, R. Fonner, M. A. Haltuch, K. Ono, and H. Winker. Accounting for spatio-temporal variation and fisher targeting when estimating abundance from multispecies fishery data. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(999):1–14, 2016.

[140] P. J. Diggle, R. Menezes, and T. L. Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 59:191–232, 2010.

[141] T. R. Carruthers, R. N. M. Ahrens, M. K. McAllister, and C. J. Walters. Integrating imputation and standardization of catch rate data in the calculation of relative abundance indices. *Fisheries Research*, 109:157–167, 2011.

[142] P. B. Conn, J. T. Thorson, and D. S. Johnson. Confronting preferential sampling in wildlife surveys: diagnosis and model-based triage. *bioRxiv*, 10.1101/080879, 2016.

[143] K. Kristensen. TMB: general random effect model builder tool inspired by ADMB. R package version 1.1., 2014.

[144] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:319–392, 2009.

[145] R. A. Webster. Results of space-time modelling of IPHC fishery-independent setline survey WPUE and NPUW data. Aailable at `https://www.iphc.washington.edu/publications/rara/2016/IPHC-2016-RARA-26-R-3.5_Results_of_space-time_modelling.pdf`, 2017.

[146] M. F. Sigler and C. R. Lunsford. Effects of individual quotas on catching efficiency and spawning potential in the Alaska sablefish fishery. *Canadian Journal of Fisheries and Aquatic Sciences*, 58:1300–1312, 2001.

[147] S. D. Hoyle and H. Okamoto. Analyses of Japanese longline operational catch and effort for bigeye and yellowfin tuna in the WCPO. western and Central Pacific Fisheries Commission. Scientific Committee, Seventh Regular Session. `http://www.spc.int/DigitalLibrary/Doc/FAME/Meetings/WCPFC/SC7/SA-IP-01.pdf`, 2011.

[148] J. Hampton, J. R. Sibert, P. Kleiber, M. N. Maunder, and S. J. Harley. Fisheries decline of Pacific tuna populations exaggerated? *Nature*, 434:E1–E2, 2005. 10.1038/nature03581.

[149] M. Ichinokawa and J. Brodziak. Using adaptive area stratification to standardize catch rates with application to North Pacific swordfish (*Xiphias gladius*). *Fisheries Research*, 106:249–260, 2010.

[150] S. Tian, Y. Chen, X. Chen, L. Xu, and X. Dai. Impacts of spatial scales of fisheries and environmental data on catch per unit effort standardisation. *Marine and Freshwater Research*, 60:1273–1284, 2010.

[151] M. Plummer. JAGS version 4.0.0 user manual. Available at `http://ftp.stu.edu.tw/pub/BSD/FreeBSD/distfiles/mcmc-jags/jags_user_manual.pdf`, 2015.

[152] D. J. Spiegelhalter, A. Thomas, N. Best, W. Gilks, and D. Lunn. BUGS: Bayesian inference using Gibbs sampling. *Version 0.5ii*, 19, 1996.

[153] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10:325–337, 2000.

[154] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. WinBUGS user manual, 2003.

[155] S. Sturtz, U. Ligges, and A. E. Gelman. R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software*, 12:1–16, 2005.

[156] J. Thompson, T. Palmer, and S. Moreno. Bayesian analysis in Stata using WinBUGS. *The Stata Journal*, 6:530–549, 2006.

[157] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS book: a practical introduction to Bayesian analysis*. CRC press, 2012.

[158] A. Thomas. OpenBUGS. `www.openbugs.net`, 2004.

[159] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: evolution, critique and future directions. *Statistics in Medicine*, 28:3049–3067, 2009.

[160] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. OpenBUGS user manual, version 3.2.3, 2014.

[161] M. B. Thompson. A comparison of methods for computing autocorrelation time. *arXiv preprint arXiv:1011.0175*, 2010.

[162] Stan Development Team. RStan: the R interface to Stan. version 2.9.0. `http://mc-stan.org/rstan.html`, 2015.

[163] M. U. Gruebler and B. Naef-Daenzer. Fitness consequences of pre- and post-fledging timing decisions in a double-brooded passerine. *Ecology*, 89:2736–2745, 2008.

Appendix A

CITATION PATTERNS OF BAYESIAN SOFTWARE PACKAGES

We analyzed the annual number of citations to the major Bayesian software packages across fields by accessing the ISI Web of Science Core Collection on 8 March 2016. Many citations were made to user manuals, books, and book chapters, and therefore citations were obtained from the Cited Reference Search. Citations to all years and versions of user manuals were included. Each citing paper was counted only once for each software package, even if it cited multiple papers describing a software package. We also excluded the year 2016 due to incomplete records.

We found $n = 5,234$ citations, after filtering, in $n = 1,377$ unique journals. The following references were included in the analysis:

Stan: [37, 29, 36]

JAGS: [38, 151]

BUGS: [152]

WinBUGS:[153, 154, 155, 156, 157]

OpenBUGS: [158, 159, 160]

Appendix B

MODEL FILES AND CODE

This study can be reproduced using free, open source software. The interested reader can find model files, simulation code, and implementation of the static HMC and NUTS algorithms, written in R, online at: `https://github.com/colemonnahan/gradmcmc`.

Appendix C

MEASURING MCMC EFFICIENCY

There are many factors contributing to which software package is 'best' for a real problem. Aspects such as learning curve, model development time, diagnostics tools, and documentation typically play a large role. Here, we assume that run time is a limiting factor preventing inference for an applied problem, so that more 'efficient' algorithms are desired. There are two key aspects of a metric that measures efficiency: run time and the minimum effective number of samples (ESS) produced.

Run time has been quantified using both log density evaluations and processor time [161], with the former being preferable when comparisons are within the same software. However, here we are comparing across software platforms, and further, Stan is also calculating gradients so model evaluations are not directly comparable. As such, we adopt processor time as the metric of run time for this study. Run time includes the warmup period but not the time to compile Stan models.

ESS is typically lower than the nominal sample size because samples drawn using MCMC are dependent (i.e., autocorrelated) and thus contain less information than the nominal sample size. ESS must be approximated from the samples [161], and to date have usually been calculated using a package like CODA [85]. However, a new formula presented in [29] is more reliable and accurate, and thus generally preferred. Although the formula is the default in Stan it is available for any MCMC output (e.g., with JAGS) by using the function monitor in the R package rstan [162], which we use to calculate ESS for all models in this study. Since each parameter has a different ESS, we take the minimum across parameters as the bottleneck of information, following the lead of other studies [55, 29].

Thus *efficiency* has a specific definition throughout this study: the estimated number of

effective samples drawn per unit time for the slowest mixing parameter.

Appendix D

EFFICIENT PARAMETERIZATIONS FOR HIERARCHICAL MODELS

Hierarchical models can be particularly challenging for MCMC samplers due to the geometry of the hypervariance and random effects. For small values of the hypervariance, each random effect will be constrained to be small. As the hypervariance increases, so will the distribution of the random effects, creating a funnel shaped relationship between the two (see Fig. 3 of [47]). The more exaggerated this geometric shape, the harder it is generally for MCMC samplers to efficiently sample from the posterior [46].

This is a problem for HMC in particular because the optimal step size will vary with the hypervariance: smaller step sizes are appropriate for smaller hypervariances and vice versa. Further, this shape cannot be corrected by a global rotation or scaling (via the mass matrix), and the slow mixing subsequently impacts the tuning of the mass matrix because the sampler will not explore the whole space during warmup [47].

Fortunately this issue can often be mitigated, at least to some degree, by reparameterizing the hyperdistribution from a *centered* into a *noncentered* form [46]. This reparameterization technique may be unfamiliar to ecologists, but can make substantial improvements to efficiencies for both HMC and other samplers as well. Noncentering works on all distributions in the location and scale family, as detailed in [36], but here we demonstrate this technique on normally distributed random effects. Let μ and σ be the hyperparameters, then let

$$
\begin{aligned}
X_1 \sim & N(\mu, \sigma^2) \\
Z \sim & N(0, 1) \\
X_2 = & \mu + \sigma Z \Rightarrow X_2 \sim N(\mu, \sigma^2)
\end{aligned}
$$

The random variable X_1 is the centered form and the natural way to parameterize the model, while X_2 the noncentered parameterization. Here both X_1 and X_2 have the same distribution but the parameters (and hence the posterior geometries) are different. In the centered form we directly model the random effects, X_1. In the noncentered form the parameters are Z, and the random effects used in the model, X_2, are derived from these.

The benefit of this is that the correlation between σ^2 (the hypervariance) and the random effects is minimized, mitigating the funnel shaped relationship between them. However, this increases the correlation between the fixed effects in the model. Further, if the data are informative about σ^2 (i.e., it has a narrow marginal posterior distribution) then the funnel is effectively truncated and the noncentered form may actually introduce more difficult correlations for the samplers. Thus, which parameterization is most efficient will depend on the situation: model structure, data, and priors [46].

Fortunately, it is typically straightforward in practice to switch between parameterizations for both Stan and JAGS, thus testing both forms as part of model development is straightforward and an approach we advocate. If the funnel shape cannot be eliminated or mitigated, HMC may have a hard time exploring all parts of the posterior. In this case, it is recommended to run a chain with smaller step sizes (i.e., set `adapt_delta` closer to 1) and verify it explores the same posterior space [47].

Appendix E

CHAPTER 1 CASE STUDY DETAILS

Here we provide further details of our case studies, summarized in Table 1.1. Note that model files for both Stan and JAGS are available at `https://github.com/colemonnahan/gradmcmc/tree/master/models`. For hierarchical models, we fit both a centered and non-centered form for the random effects, with the latter having names appended with '_nc' to distinguish them.

E.1 MVND and MVNC

Our first simulation model is a multivariate normal. MCMC methods are not needed here, but this model provides a convenient testing framework because it is easy to scale dimensionality and complexity (via the covariance structure). We use this model in two ways. First, we generate a 50 by 50 covariance matrix with a mixture of weak and strong correlations using an inverse Wishart distribution (with 50 df and diagonal covariance), as done in [29], varying the dimensionality by taking subsets of this matrix (MVND) from 2 to 50. For each subset we also run a diagonal covariance matrix (i.e., independent) to differentiate the effect of model size vs. complexity (correlations).

Second, we generate a covariance matrix with constant correlations ($0 \leq \rho \leq 0.95$) for all off-diagonal entries. We varied ρ from 0 (independent) to 0.95, and for each level of ρ we run the model for 2, 25, and 50 dimensions, again to differentiate effects of size and correlation. We name this model MVNC.

Samplers which update a single parameter at a time (e.g., Gibbs and slice sampling, typically) are known to be sensitive to correlations or 'scale invariant.' Conversely, HMC, which updates all parameters simultaneously, is known to be sensitive to variable scales

but not correlations, and is thus 'rotation invariant.' However, if the mass matrix closely matches the global covariance structure (and it is constant across the surface) then HMC will be scaled appropriately [34]. These differences motivate our decision to test both MVND, which has different scales, and MVNC which has the same scale but stronger correlations.

For both these models we used broad, non-conjugate uniform priors, so that Gibbs sampling was not possible and note that slice sampling was utilized by JAGS. These models mimic fixed effects models with non-conjugate priors.

E.2 Growth

The second simulation model, 'Growth,' is a non-linear somatic growth model with repeated measures [48]. This model has two sets of random effects (two measurements for each individual), and fixed effects for hyperparameters and a term controlling the shape of the growth pattern.

We simulated data by drawing from independent hyperdistributions for k (growth rate) and $L^{\max}$ (mean maximum length), both on the log scale, sampling at 5 random ages, and then adding normal measurement error to length on the log scale. This simulates resampling the same individual multiple times over its lifetime, as would be done for example in a controlled lab study. Our estimation model matched the structure of the simulated data, so that each animal had a corresponding random effect for $\log k$ and $\log L^{\max}$, as well as fixed effects for hypermeans, hypervariances, an observation variance term for the normal likelihood.

Similar to the multivariate normal examples, this model let us explore the effect of dimensionality (by increasing the number of individuals), and how that interacted with noncentering the random effects. We ran the model with number of individuals from 2 to 316. We used an `adapt_delta` of 0.8 for both the centered and noncentered forms, although higher values would be necessary to minimize divergent transitions for models with few individuals in practice. In this case, we favored a constant target acceptance rate to more easily compare performance and test for divergence.

The estimation model is:

$$\begin{aligned}
L_{i,j} &= i\text{th observed length of animal } j \text{ at age } a_{i,j} \\
\hat{L_{i,j}} &= L_j^{\max}\left(1 - exp(-k_j(a_{i,j} - t_0))\right)^{\delta} \\
\log L_j^{\max} &\sim N(\mu_{L^{\max}}, \sigma_{L^{\max}}) \\
\log k_j &\sim N(\mu_k, \sigma_k) \\
L_{i,j} &\sim N(\hat{L}_{i,j}, \sigma_{\text{obs}})
\end{aligned}$$

Priors were specified as:

$$\begin{aligned}
\delta &\sim U(0, 5) \\
\mu_k &\sim U(-5, 5) \\
\sigma_k &\sim \text{Half-Cauchy}(0, 25) \\
\mu_{L^{\max}} &\sim U(-5, 5) \\
\sigma_{L^{\max}} &\sim \text{Half-Cauchy}(0, 25) \\
\sigma_{\text{obs}} &\sim \text{Half-Cauchy}(0, 25)
\end{aligned}$$

Simulated values were:

$$\begin{aligned}
\mu_k &= -2.3 \\
\sigma_k &= 0.2 \\
\mu_L^{\max} &= 3.9 \\
\sigma_L^{\max} &= 0.1 \\
\sigma_{\text{obs}} &= 0.1 \\
t_0 &= 5
\end{aligned}$$

E.3 Redkite

This model is described in detail in section 8.4 of [49]. The data come from marking 1480 nestlings and 152 adults over the previous 50 years. The model estimates age-specific survival probabilities of three age classes: juveniles, subadults and adults. This model contains an informative prior on juvenile survival, because it is often difficult to estimate with these kinds of data.

This model was converted to Stan and posted online at `https://github.com/stan-dev/example-models/tree/master/BPA` (last accessed 5/8/2016), while the original BUGS code was taken from `http://www.vogelwarte.ch/de/projekte/publikationen/bpa/` (last accessed 5/8/2016).

E.4 Swallows

The 'Swallows' model estimates survival of birds accounting for survival and detection using a state-space construction, while linking survival and detection probabilities to environmental covariates. It also includes a complex hierarchical design, with year and family random effects for survival, and family random effects for detection [163, 50].

The original model code can be found online at `http://www.oikostat.ch/blmeco/ch14/CJS_swallows.stan` (last accessed 5/4/2016). We modified this version to include recommended priors for the hyper-variance terms, additional data preprocessing, and a few computational improvements (such as vectorization). [50] found that their BUGS implementation was prohibitively slow compared to Stan. However, that implementation tracked discrete states for each animal in each time step. That is, it did not marginalize out the discrete parameters. When we recreated the model in JAGS we used the same parameterization as Stan. Thus it seems likely that the major increase in efficiency that we found was due to the parameterization, not the actual software platform or algorithm.

E.5 Logistic

This model is a state-space population dynamics model with a logistic population growth function applied to South Atlantic tuna catch and effort data. Process errors, observation variance, and biological parameters are all estimated, some of which have informative priors derived from meta-analyses. This model is described in detail in [16], including the original BUGS code. We used `adapt_delta` of 0.95 for both forms to minimize divergences.

E.6 Wildflower

This model was analyzed in [32], and looks at individual flowering success of Bitterroot milkvetch (*Astragulus scaphoides*) as a function of the last year's stage and seed production. It is a binomial generalized linear mixed effects model, with three sets of random effects, two of which are crossed. The first set acts as an intercept on year, while the latter two are individual effects for slope and intercept. The fixed effects act as intercept on stage, a global slope, and hypervariances.

Following the original authors, we did not center the predictor variables. However, we did introduce the recommended half-Cauchy priors for the variance terms, instead of bounding their values and using a uniform prior as originally done.