2017

Improving the Computational Efficiency in Bayesian Fitting of Cormack-Jolly-Seber Models with Individual, Continuous, Time-Varying Covariates

Woodrow Burchett
University of Kentucky, woodrow.burchett@uky.edu

Digital Object Identifier: https://doi.org/10.13023/ETD.2017.250

Woodrow Burchett, Student

Dr. Simon Bonner, Major Professor

Dr. Constance Wood, Director of Graduate Studies

Improving the Computational Efficiency in Bayesian Fitting of Cormack-Jolly-Seber Models with Individual, Continuous, Time-Varying Covariates

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Woodrow Burchett
Lexington, Kentucky

Co-Directors: Dr. Simon Bonner, PhD, Professor of Statistics
and Dr. Arnold Stromberg, PhD, Professor of Statistics
Lexington, Kentucky 2017

ABSTRACT OF DISSERTATION

Improving the Computational Efficiency in Bayesian Fitting of Cormack-Jolly-Seber Models with Individual, Continuous, Time-Varying Covariates

The extension of the CJS model to include individual, continuous, time-varying covariates relies on the estimation of covariate values on occasions on which individuals were not captured. Fitting this model in a Bayesian framework typically involves the implementation of a Markov chain Monte Carlo (MCMC) algorithm, such as a Gibbs sampler, to sample from the posterior distribution. For large data sets with many missing covariate values that must be estimated, this creates a computational issue, as each iteration of the MCMC algorithm requires sampling from the full conditional distributions of each missing covariate value. This dissertation examines two solutions to address this problem. First, I explore variational Bayesian algorithms, which derive inference from an approximation to the posterior distribution that can be fit quickly in many complex problems. Second, I consider an alternative approximation to the posterior distribution derived by truncating the individual capture histories in order to reduce the number of missing covariates that must be updated during the MCMC sampling algorithm. In both cases, the increased computational efficiency comes at the cost of producing approximate inferences. The variational Bayesian algorithms generally do not estimate the posterior variance very accurately and do not directly address the issues with estimating many missing covariate values. Meanwhile, the truncated CJS model provides a more significant improvement in computational efficiency while inflating the posterior variance as a result of discarding some of the data. Both approaches are evaluated via simulation studies and a large mark-recapture data set consisting of cliff swallow weights and capture histories.

KEYWORDS: Mark-recapture; Bayesian Inference; Variational Bayes; Individual time-varying continuous covariates

Author's signature: Woodrow Burchett

Date: June 27, 2017

Improving the Computational Efficiency in Bayesian Fitting of Cormack-Jolly-Seber Models with Individual, Continuous, Time-Varying Covariates

By

Woodrow Burchett

Co-Director of Dissertation: Simon Bonner, PhD

Co-Director of Dissertation: Arnold Stromberg, PhD

Director of Graduate Studies: Constance Wood, PhD

Date: June 27, 2017

ACKNOWLEDGMENTS

This dissertation would not have been possible without the leadership, guidance and constant support of my advisor, Dr. Simon Bonner. Dr. Bonner's patience and enthusiasm are seemingly without limit and I cannot possibly thank him enough for his assistance.

I would like to express gratitude to Dr. Arnold Stromberg, Dr. Katherine Thompson, Dr. William Griffith, and Dr. David Westneat for serving on my committee and offering valuable suggestions and insights during this process. Thanks also to Dr. Kwok-Wai Ng for serving as the outside examiner.

I would also like to express my thanks to Dr. Matthew Schofield for initiating this project. Dr. Schofield's enthusiasm, advice and direction were integral to the development of this dissertation.

Lastly, I would like to thank my friends and family for their support and encouragement throughout this endeavor.

TABLE OF CONTENTS

Acknowledgments iii

Table of Contents iv

List of Figures vi

List of Tables ix

Chapter 1 Introduction 1

1.1 Overview 1

1.2 Mark Recapture Methods 6

1.3 Cormack-Jolly-Seber Model 8

1.4 Continuous Covariates 10

1.5 Cliff Swallows Data 13

Chapter 2 Variational Bayes 15

2.1 Introduction to Variational Bayesian Methods 15

2.2 Application of the Mean Field Approach to the CJS Model with Continuous Covariates 19

Chapter 3 Improvements to the Variational Bayesian Algorithm 37

3.1 Mixed Effects Model 37

3.2 Modified Mean Field Approach to the CJS Model with Continuous Covariates 44

3.3 Hybrid Algorithm 54

3.4 Application to Cliff Swallows 60

3.5 Discussion 63

Chapter 4 Truncated CJS 66

4.1 Introduction to Truncated CJS Model 66

4.2 Truncated CJS Likelihood 68

4.3 Accuracy and Precision of the Truncated CJS Model 69

4.4 Simulation Study 73

4.5 Application to Cliff Swallows 79

4.6 Discussion 85

Chapter 5 Conclusion 88

Chapter A Appendices 92

A.1 Proof of Mean Field Maximization Result 92

A.2 Expected Values for MFVB Algorithm 95

A.3 Derivation of Optimal Variational Densities for the MFVB Method Applied to the Mixed Effects Model 97

A.4 Temporary Emigration 99

Bibliography 113

Vita 117

LIST OF FIGURES

2.1 Iterations Required for Convergence: MFVB vs MCMC 29

2.2 MFVB and MCMC Parameter Estimates in Low Capture Scenario . . . 31

2.3 Simulation Results Comparing MFVB and MCMC in Low Capture Scenario 32

2.4 MFVB and MCMC Parameter Estimates in High Capture Scenario . . . 34

2.5 Simulation Results Comparing MFVB and MCMC in High Capture Scenario 35

3.1 Ormerod and Wand's Algorithm and MCMC Parameter Estimates from Orthodont Data 39

3.2 Ormerod and Wand's Algorithm and MCMC Parameter Estimates from Simulated Data 40

3.3 New Algorithm, Ormerod and Wand's Algorithm, and MCMC Parameter Estimates from Orthodont Data 43

3.4 New Algorithm, Ormerod and Wand's Algorithm, and MCMC Parameter Estimates from Simulated Data 44

3.5 Iterations Required for Convergence: MFVB vs Corrected MFVB vs MCMC 49

3.6 MFVB, Corrected MFVB, and MCMC Parameter Estimates in Low Capture Scenario 51

3.7 Simulation Results Comparing MFVB, Corrected MFVB, and MCMC in Low Capture Scenario 52

3.8 MFVB, Corrected MFVB, and MCMC Parameter Estimates in High Capture Scenario 53

3.9 Simulation Results Comparing MFVB, Corrected MFVB, and MCMC in High Capture Scenario 54

3.10 MFVB, Corrected MFVB, Hybrid MFVB, and MCMC Parameter Estimates in Low Capture Scenario 56

3.11 MFVB, Corrected MFVB, Hybrid MFVB, and MCMC Parameter Estimates in High Capture Scenario 57

3.12 Simulation Results Comparing MFVB, Corrected MFVB, Hybrid MFVB, and MCMC in Low Capture Scenario 58

3.13 Simulation Results Comparing MFVB, Corrected MFVB, Hybrid MFVB, and MCMC in High Capture Scenario 59

3.14 Comparison of Parameter Estimates for a Subset of Cliff Swallows Data . 61

3.15 Comparison of Parameter Estimates for Cliff Swallows Data 62

4.1 KL Distances between MLEs at Different Values of k and p 72
4.2 KL Distances between MLEs at Different Values of k and ϕ 73
4.3 Parameter Estimates from the Truncated CJS Model at Different Values of k for Simulated Data with $T = 15$ 77
4.4 Parameter Estimates from the Truncated CJS Model at Different Values of k for Simulated Data with $T = 20$ 79
4.5 Estimates of Capture Parameters from the Truncated CJS Model for 8 Years of Cliff Swallows Data 81
4.6 Estimates of Survival Parameters from the Truncated CJS Model for 8 Years of Cliff Swallows Data 82
4.7 Estimates of Capture Parameters from the Truncated CJS Model for Cliff Swallows Data 83
4.8 Estimates of Survival Parameters from the Truncated CJS Model for Cliff Swallows Data 84

A.1 Estimates of Capture Parameters from the Truncated CJS Model for a Single Simulated Data Set with No Temporary Emigration 100
A.2 Estimates of Survival Parameters from the Truncated CJS Model for a Single Simulated Data Set with No Temporary Emigration 101
A.3 Estimates of Capture Parameters from the Truncated CJS Model for a Single Simulated Data Set with Severe Temporary Emigration 102
A.4 Estimates of Survival Parameters from the Truncated CJS Model for a Single Simulated Data Set with Severe Temporary Emigration 103
A.5 Estimates of Capture Parameters from the Truncated CJS Model for a Single Simulated Data Set with Moderate Temporary Emigration 104
A.6 Estimates of Survival Parameters from the Truncated CJS Model for a Single Simulated Data Set with Moderate Temporary Emigration 105
A.7 Average Estimates of Capture Parameters from the Truncated CJS Model across 100 Simulated Data Sets with No Temporary Emigration 106
A.8 Average Estimates of Survival Parameters from the Truncated CJS Model across 100 Simulated Data Sets with No Temporary Emigration 107
A.9 Average Estimates of Capture Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Severe Temporary Emigration . . . 108
A.10 Average Estimates of Survival Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Severe Temporary Emigration . . . 109

A.11 Average Estimates of Capture Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Moderate Temporary Emigration . 110

A.12 Average Estimates of Survival Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Moderate Temporary Emigration . 111

LIST OF TABLES

1.1 Possible probabilities assigned to an individual's capture history 10

3.1 Differences between the original MFVB algorithm and the correlation corrected MFVB algorithm 48

4.1 Possible probabilities assigned to individual i's capture history 68

4.2 Run times, estimated % bias, average relative standard errors, and effective samples per second when fitting the truncated CJS model (at 4 different values of k) to a data set with $n = 600$ and $T = 15$ capture occasions. 76

4.3 Run times, estimated % bias, average relative standard errors, and effective samples per second when fitting the truncated CJS model (at 4 different values of k) to a data set with $n = 600$ and $T = 20$ capture occasions. 78

4.4 Run times and effective samples per second when fitting the truncated CJS model (at 4 different values of k) to the first 8 years of the cliff swallow data. 82

Chapter 1 Introduction

1.1 Overview

Mark-recapture studies have been performed by ecological researchers for over a century, beginning with Danish biologist C.G. Johannes Petersen's study of the European plaice population in 1896 (Petersen, 1896). Since then, researchers have studied many different animal populations via mark-recapture methods, from wild Soay sheep living on an isolated Scottish island (Clutton-Brock and Pemberton, 2004) to cliff swallows nesting under bridges in the Western United States (Brown and Brown, 1996). These mark-recapture studies, which consist of capturing the animals, assigning them unique marks, and releasing them back into the population over multiple discrete capture occasions, can produce estimates of a variety of different parameters that describe the population of interest, including population size, birth rates, and survival rates. The development of methods to analyze data generated from these studies is an active area of research, incorporating many modern statistical and computational techniques such as Markov Chain Monte Carlo (Bonner and Schwarz, 2006), multiple imputation (Worthington et al., 2015), Bayesian model selection (King et al., 2008), and the expectation-maximization (EM) algorithm (Xi et al., 2009).

Many experiments are designed to estimate the survival of individuals in a population and to identify the characteristics of the animals or environment which might impact an individual's survival. Most models for studying survival from mark-recapture data are based on the Cormack-Jolly-Seber (CJS) model (Cormack, 1964; Jolly, 1965; Seber, 1965). The CJS model assigns probabilities to the individual capture histories as a function of two sets of parameters: the capture probabilities (the probability that an individual alive on a particular capture occasion is caught on that occasion) and the survival probabilities (the probability that an individual alive on a particular capture occasion will also be alive on the next capture occasion). In the original model, these probabilities are allowed to vary over time, but not between individuals in the population. Additionally, the CJS model assumes that each capture occasion occurs instantaneously, marks are not mistakenly identified or lost, death or emigration is permanent, and individuals in the population behave independently of one another (Seber, 2002, page 196).

The desire for researchers to study the effects of different factors on survival and to account for possible differences in capture susceptibility led to extensions of the CJS model being developed (Pollock, 2002). These extensions initially included models to allow for the effects of individual and environmental continuous covariates on survival and/or capture via a link function (Lebreton et al., 1992), and later those incorporating dichotomous and categorical individual covariates that change over time by way of the multi-state model (Brownie et al., 1993; Schwarz et al., 1993). The multi-state model was developed because individual, time-varying covariates are missing on capture occasions where individuals are not observed, and these missing covariate values are necessary to define the survival and/or capture probabilities. This missing data problem is solved by assuming that the covariate values follow the Markov property, which allows for the summation over all possible covariate values in the likelihood. Applying the multi-state approach to model the effect of an individual, time-varying, continuous covariate, however, would require integrating, rather that summing, over all possible missing covariate values, which results in an intractable likelihood function.

One solution to the problem of missing individual, time-varying, continuous covariates is to discretize any such variables into discrete bins and analyze the data via the multi-state model (Nichols et al., 1992). Simple imputation techniques, such as carrying the last observation forward or taking the mean of an individual's observed values, can also be implemented to address the issue of missing covariates. However, the parameter estimates can be very biased if the imputation algorithm does not closely match the true underlying process and these imputation methods do not account for the uncertainty in covariate estimation, which will lead to artificially low standard errors (Bonner et al., 2010).

Bonner and Schwarz (2006) solved this problem by explicitly modeling the missing covariates to construct a complete data likelihood. The fitting then occurs via a Bayesian framework in which Markov Chain Monte Carlo (MCMC) algorithms are applied to generate samples from the joint posterior distribution of both the model parameters and missing covariate values. This avoids the analytically intractable integration over all possible covariate values that a maximum likelihood approach would necessitate. The specific covariate model introduced in Bonner and Schwarz (2006) assumes that the differences in covariate values from one capture occasion to the next are normally distributed, with a constant precision across all capture occasions and individuals, and that the average change between subsequent capture occasions is constant for all individuals. This is a Markov process, as was the covariate model assumed in the multi-state model. King et al. (2008) applied this technique to analyze Soay sheep data, slightly modifying the Bonner and Schwarz (2006) covariate model and conducting Bayesian model selection via reversible jump MCMC to probabilistically evaluate the appropriateness of different modeling assumptions. This demonstrates that the Bayesian approach to the estimation of missing continuous covariates is both flexible, allowing the model for the missing covariates to be easily modified to fit different real world problems, and that the assumptions made in the modeling process can be rigorously examined via Bayesian model selection techniques.

The main assumptions underlying the model for the missing covariates presented by Bonner and Schwarz (2006) are that the change in an individual's covariate value on consecutive capture occasions is normally distributed, that the mean change between subsequent capture occasions is constant across all individuals in the population, that the variance or precision of this process is constant across sampling occasions and individuals, and that the changes are independent across sampling occasions and between individuals.

This approach, however, has limitations when it comes to analyzing very large data sets. Markov Chain Monte Carlo methods that rely on sampling repeatedly from full conditional distributions until convergence is reached often scale very poorly as the sample size increases, especially when every missing covariate value must be generated on each iteration of the algorithm. This repeated sampling of a potentially large number of missing covariates can make the traditional MCMC approach unfeasible, especially when the experiment was conducted over many capture occasions and individuals are short lived and/or capture rates are low so that there are many occasions where individuals are not captured.

Langrock et al. (2013) attempted to address these issues by returning to the maximum likelihood framework. In particular, they finely discretized the continuous covariates to facilitate numerically integrating over the range of possible values, extending the coarse binning approach found in Nichols et al. (1992). The resulting likelihood is equivalent to that of a hidden Markov model, which allowed the authors to take advantage of an efficient, recursion-based evaluation of the likelihood function. This increased efficiency in evaluating the likelihood makes maximum likelihood estimation feasible. Additionally, although the discretization of the continuous covariate results in the maximization of an approximate likelihood, this approximation can be made arbitrarily more accurate by more finely discretizing the covariate at the cost of computational efficiency. Langrock et al. (2013) mentioned that in the presence of two continuous, time-varying individual covariates, however, the Bayesian approach introduced by Bonner and Schwarz (2006) may be preferable, as the computational burden for their maximum likelihood method quickly becomes untenable. Additionally, although Langrock et al. (2013) specifically considered mark-recapture-recovery data, the technique they introduce also applies to mark-recapture data. Note that a mark-recapture-recovery experiment is simply a mark-recapture study in which deceased individuals may be recovered. Also note that although the model proposed by Bonner and Schwarz (2006) was originally fit to mark-recapture data, it could just as easily be applied to mark-recapture-recovery data.

Worthington et al. (2015) approached the problem by applying the technique of multiple imputation to facilitate maximum likelihood estimation. This method begins by first modeling only the continuous covariates and then generating multiple complete sets of covariates from the fitted model. Maximum likelihood estimates can then be obtained very quickly by fitting the CJS model to each of the generated data sets with complete covariate information. This set of estimates can then be aggregated via non-parametric bootstrap techniques in order to appropriately account for the uncertainty in the estimation of the covariates. In addition to being significantly faster than the Bayesian approach introduced by Bonner and Schwarz (2006), it also avoids the computational issues present in Langrock et al. (2013) when incorporating multiple continuous covariates. The downside to the multiple imputation approach is that it relies on the assumption that the covariates are missing at random (i.e. the information contained in the capture histories is ignored when imputing the missing covariates). The authors admit that this is an unrealistic assumption, and while this method performs extremely well in the simulation results presented in Worthington et al. (2015), the simulation study only considered a covariate effect on survival. If there was a covariate effect on the capture probabilities, then the missing at random assumption would be more severely violated and I believe that substantial bias in the parameter estimates could occur.

Another approach to estimate the effects of continuous covariates on survival probabilities estimated from mark-recapture-recovery data was introduced by Catchpole et al. (2008). This method produces parameter estimates without the need to impute or model any missing covariates and is known as the trinomial model. The trinomial model only considers events on the occasions directly following capture occasions on which an individual was captured and the covariate measured. The likelihood, containing only information from those capture occasions, is then maximized to obtain parameter estimates without the need to estimate or impute any missing covariates or assume any model associated with the covariates, as all of the survival probabilities included in the likelihood will have an associated observed covariate. One drawback of this method is the increased variance of the parameter estimates, as potentially useful information contained in individual capture histories is discarded when there is no available covariate information. This model also relies heavily on the recovery of deceased individuals present in mark-recapture-recovery data and has difficulties when covariates are associated with capture probabilities. Bonner et al. (2010) provided a thorough comparison of the Bayesian missing covariate estimation and the trinomial model introduced by Catchpole et al. (2008) and found that the trinomial method can produce biased results when capture probabilities and/or sample sizes are low.

Bonner (2003) explored the implementation of an EM algorithm to address the missing data challenge when fitting the CJS model with individual, continuous, time-varying covariates. Unfortunately, the expectation step of the algorithm requires the evaluation of multi-dimensional integrals with no analytic solutions. Bonner (2003) attempted to solve this issue by approximating the expected values numerically through Monte Carlo integration. However, the variability associated with the parameter estimates needed to be bootstrapped. This made the method extremely computationally demanding, as Monte Carlo integration needed to occur on every iteration of the EM algorithm which itself needed to be run multiple times to generate bootstrapped variability estimates and confidence intervals. Additionally, the Monte Carlo EM approach did not perform as well as a Bayesian MCMC algorithm in simulation studies. Xi et al. (2009), while not directly addressing this problem, did successfully implement an EM algorithm to solve a missing data problem in the case of a closed population model (i.e. individuals cannot die or leave the population throughout the duration of the study) where the covariates are not time-varying.

In this dissertation, I attempt to solve the missing, individual, time-varying, continuous covariate problem by exploring two very different approaches. The first involves abandoning the MCMC methodology for sampling from the posterior distribution in favor of analytical approximation (specifically, variational Bayesian techniques). Variational Bayesian methods provide an alternative to MCMC algorithms and have been widely applied in the field of computer science (Jordan et al., 1999; Jaakkola and Jordan, 2000; Minka, 2001; Mandt and Blei, 2014; Polatkan et al., 2015). These techniques are significantly faster and deterministic, but rely on making some assumptions about the posterior distributions to simplify the estimation process (Jordan et al., 1999; Ormerod and Wand, 2010). The idea behind variational Bayesian methodology is to replace the potentially time consuming and resource intensive sampling that occurs in an MCMC algorithm with an optimization problem that is rendered tractable by making some assumptions about the underlying posterior distribution. The increase in speed comes at the cost of deriving inference from an approximate posterior distribution (rather than sampling from the true posterior distribution, as MCMC does) restricted by the aforementioned assumptions. Additionally, this approximate posterior distribution usually underestimates the variability of the true posterior distribution (Ormerod and Wand, 2010).

My second approach relies on a different approximation to the posterior distribution obtained by altering the CJS likelihood to allow the truncation of capture histories, the same basic idea underlying the method described in Catchpole et al. (2008). This will reduce the number of missing covariates that must be imputed by focusing on the missing data that has the most influence over the parameter estimates. To accomplish this, I truncate individual capture histories after each recapture according to a tuning parameter k. I call this approach the truncated CJS model. This method does discard some data and, as a result, produces posterior samples with higher variance than that of the true posterior distribution. As I will show, however, the posterior estimates produced when fitting this model are still unbiased and carefully choosing the value of k can result in an MCMC algorithm capable of generating samples from a posterior distribution that are almost indistinguishable from samples of the true posterior distribution in a fraction of the time.

The manuscript begins with an introduction to mark-recapture methods in Section 1.2, followed by the definition of the original CJS model in Section 1.3, the extension to time-varying, individual, continuous covariates in Section 1.4, and a description of the large mark-recapture data set I will analyze as an example when evaluating both of my new methods in Section 1.5. I then apply a standard variational Bayesian approach to the CJS model with individual, time-varying, continuous covariates in Chapter 2 and describe some of issues with the approximation. In Chapter 3 I produce a better variational Bayesian approximation at the cost of a large computational burden and introduce a method that combines this more accurate approximation with the faster algorithm from Chapter 2. In Chapter 4 I take a different approach and introduce the truncated CJS model, which I fit using MCMC algorithms. Finally, I conclude with some discussion about my findings in Chapter 5.

1.2 Mark Recapture Methods

Before any discussion of models or fitting algorithms can begin, I must first define the structure of data collected during mark-recapture studies. I begin by providing a list of notation. Then, I describe how data is collected and recorded during a mark-recapture study.

Notation

The following list describes the data and parameters necessary to define the CJS model:

- Study Parameters

$$T = \text{number of capture occasions}$$

- Observed Data

$$\begin{aligned}
n &= \text{number of individuals marked during the } T \text{ capture occasions}\\
\omega_{i,t} &= \begin{cases} 1, & \text{if individual } i \text{ is captured at occasion } t\\ 0, & \text{otherwise}\end{cases}\\
\boldsymbol{\omega}_i &= (\omega_{i,1}, \omega_{i,2}, \ldots, \omega_{i,T}) = \text{capture history for individual } i\\
\Omega &= n \text{ by } T \text{ matrix where the } i\text{th row is } \boldsymbol{\omega}_i\\
a_i &= \text{first capture occasion on which individual } i \text{ was captured}
\end{aligned}$$

- Model Parameters

$$\begin{aligned}
p_t &= \text{probability that an individual is captured on}\\
&\quad \text{sampling occasion } t\text{, given that the individual is alive}\\
\phi_t &= \text{probability that an individual survives to occasion}\\
&\quad t+1\text{, given that the individual is alive at occasion } t\\
\chi_t &= \text{probability that an individual is not observed after occasion } t,\\
&\quad \text{given that the individual was captured alive on occasion } t
\end{aligned}$$

Mark Recapture Data

A mark-recapture study proceeds by sampling individuals from a population over T distinct capture occasions. This process begins on the first capture occasion, when individuals are captured, given unique marks, and released back into the population.

On subsequent capture occasions, both marked and unmarked individuals are captured. On these subsequent capture occasions, the presence of previously marked individuals is noted and unmarked individuals are given a unique mark. Both previously marked and unmarked individuals are then released back into the population.

At the conclusion of a mark-recapture study, n unique individuals have been recorded across T capture occasions. The capture histories of these individuals, after being recorded on each capture occasion, are stored in an n by T matrix $\mathbf{\Omega}$. The i,t-th entry in this matrix, $\omega_{i,t}$ is an indicator variable that takes the value 1 if individual i was captured on occasion t and 0 if not. The ith row of $\mathbf{\Omega}$, $\boldsymbol{\omega}_i = (\omega_{i,1}, \ldots, \omega_{i,T})$, is defined as the capture history for individual i. For example, suppose that I have data from a mark-recapture study with $T = 3$ capture occasions and that individual i was first captured on the 1st capture occasion and later captured on the 3rd capture occasion. That individual's capture history would be $\boldsymbol{\omega}_i = (101)$.

In addition, it is common for covariates of interest to be recorded when individuals are captured. These covariates may be static, such as gender, and need only be recorded on an individual's initial capture while others may be time varying, such as size or weight, and must be recorded on each occasion on which an individual is captured. Furthermore, these covariates can often be environmental and not related to individuals at all, such as temperature or rainfall. In the next section, I describe the original formulation of the CJS model, which does not allow for the effect of covariates. Later in Section 1.4, however, I describe the extension of the CJS model to include covariates.

1.3 Cormack-Jolly-Seber Model

Fitting a model to the mark recapture data described in the previous section allows researchers to estimate parameters of interest about the population from which the individuals are sampled. The basis for most models of open population mark-recapture studies is the CJS model (Cormack, 1964; Jolly, 1965; Seber, 1965) which, in its original form, assigns probabilities to the individual capture histories as a function of two sets of parameters: capture probabilities (the probability that an individual alive on a particular capture occasion is caught on that occasion) and survival probabilities (the probability that an individual alive on a particular capture occasion will also be alive on the next capture occasion). The CJS model assigns these probabilities to capture histories conditional on each individual's first release and assumes that individuals behave independently, sampling occasions are instantaneous, marks are not lost or overlooked, and all death or emigration from the population is permanent. See Seber (2002, page 196) for more details on the assumptions associated with the CJS model.

To define the likelihood associated with the CJS model, I must first define some notation. Let p_t represent the probability that an individual will be captured on sampling occasion t (given that the individual is alive on occasion t), ϕ_t represent the probability that an individual will survive until occasion $t+1$ (given that the individual is alive on occasion t), and χ_t represent the probability that an individual is not observed after occasion t (given that the individual is captured alive on occasion t). Note that χ_t is a function of $\mathbf{p}$ and $\boldsymbol{\phi}$, where $\mathbf{p} = (p_1, p_2, \ldots, p_t)$ and $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_t)$, and can be defined recursively as:

$$\chi_t = (1 - \phi_t) + \phi_t(1 - p_{t+1})\chi_{t+1},\ t = 1, ..., T-1$$
$$\text{and } \chi_T = 1.$$

Probabilities which are functions of $\mathbf{p}$, $\boldsymbol{\phi}$, and the derived quantity χ may then be assigned to each capture history.

As an example, consider the hypothetical individual mentioned in the previous section with a capture history of $\boldsymbol{\omega} = (101)$. This individual's capture history would be assigned the probability $\phi_1(1 - p_2)\phi_2 p_3$. The survival parameters ϕ_1 and ϕ_2 are included because I know that this individual survived from capture occasion 1 to capture occasion 3. Additionally, since I know this individual was alive on capture occasions 2 and 3, I can include $(1 - p_2)$ for the capture occasion on which this individual was not captured and p_3 for the capture occasion on which this individual was captured. Note that p_1 is not included, as the probability assigned to this individual's capture history is conditional on the first capture.

Table 1.1 provides a complete list of all individual capture histories and the assigned probabilities for a study with $T = 3$ capture occasions. The capture histories $\boldsymbol{\omega}_i = 001$ and $\boldsymbol{\omega}_i = 000$ are not included in the table because they do not contribute any information to the likelihood function due to the fact that probabilities assigned to capture histories under the CJS model condition on an individual's first capture.

Once probabilities are assigned to every possible capture history, the likelihood for the CJS model can be written as

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\phi}|\boldsymbol{\Omega}) = \prod_{i=1}^{n} Pr(\boldsymbol{\omega}_i|a_i) \tag{1.1}$$

where a_i denotes the occasion on which individual i was first captured. From here, the model can be easily fit to the data using maximum likelihood estimation. This fitting can also be made extremely efficient by modeling the data in terms of sufficient statistics contained in the m-array (Burnham, 1987). The m-array summarizes the data by reporting how many individuals were captured and released at each occasion, in addition to how many of these individuals were recaptured at each of the subsequent occasions. This information (a vector of size $T-1$ containing the amounts of released individuals and a $T-1$ by $T-1$ upper triangular matrix of first subsequent recaptures) is all that is required to fit the original CJS model. However, the m-array is insufficient for including individual covariates, and I will therefore consider the likelihood associated with individual capture histories for the remainder of this manuscript.

Table 1.1: *Possible probabilities assigned to an individual's capture history*

Capture History	Probability
111	$\phi_1 p_2 \phi_2 p_3$
110	$\phi_1 p_2 \chi_2$
101	$\phi_1(1-p_2)\phi_2 p_3$
100	χ_1
011	$\phi_2 p_3$
010	χ_2

1.4 Continuous Covariates

Researchers are often interested in the effects of covariates on the parameters associated with members of a population. Consider, for example, the effect of weight on the survivability of wild Soay sheep living on the Scottish island of Hirta (Clutton-Brock and Pemberton, 2004; King et al., 2008; Bonner et al., 2010). The original form of the CJS model described in Section 1.2 only allows capture and survival probabilities to vary by capture occasion, which facilitates the modeling of temporal changes in these parameters but cannot incorporate the effects of covariates. Fortunately, this model was later extended to incorporate survival and capture probabilities that are functions of covariates. Originally, these covariates were either static, such as an individual's gender, or common to all individuals, like the amount of rainfall observed before each capture occasion (Lebreton et al., 1992). Later, this model was extended to allow individual, categorical covariates via the multi-state model (Brownie et al., 1993; Schwarz et al., 1993). Bonner and Schwarz (2006) described a further extension of the CJS model in which capture and survival probabilities may be functions of continuous, time-varying, individual covariates. Fitting this model does present some challenges not present in models that include static or common covariates, however, as the values of such covariates cannot be observed on occasions on which individuals are not captured.

To solve this problem, Bonner and Schwarz (2006) modeled the distribution of the missing covariates to construct a complete data likelihood. Let $z_{i,t}$ represent the covariate for individual i at time t (missing if $\omega_{i,t} = 0$). Bonner and Schwarz (2006) considered this specific model for the continuous covariates:

$$[z_{i,t}|z_{i,t-1}, \Delta_t, \tau] \sim \mathcal{N}\left(z_{i,t-1} + \Delta_t, \frac{1}{\tau}\right) \tag{1.2}$$

where Δ_t represents the average change in the covariate from capture occasion $t-1$ to t and τ represents the precision of the change in an individual's covariate value between consecutive capture occasions. The main assumptions underlying this model for the missing covariates are that the change in an individual's covariate value on consecutive capture occasions is normally distributed, that the mean change between subsequent capture occasions is constant across all individuals in the population, that the variance or precision of this process is constant across sampling occasions and individuals, and that all individual's in the population are independent of each other.

Once the model for the covariate is defined, a link function, usually the logit, relates the covariate information to the capture or survival probabilities. For example, if I wanted to model the effect of a continuous, individual time-varying covariate on survival, then I would define the survival probabilities as:

$$\text{logit}(\phi_{i,t}) = \beta_0 + \beta_1 z_{i,t}$$

where $\phi_{i,t}$ now represents the probability that individual i survives from capture occasion t to $t+1$, given that individual i was alive on capture occasion t. Likewise, if the capture probabilities were dependent on an individual, time-varying covariate, the probability that individual i is captured on capture occasion t, given that individual i is alive on capture occasion t, would be denoted by $p_{i,t}$. Additionally, if either the capture or survival probabilities are modeled with respect to an individual, time-varying covariate, the probability that individual i is not observed after occasion t will be denoted $\chi_{i,t}$. The capture histories are then assigned probabilities which are nearly identical to those presented in Table 1.1 (for a study with 3 capture occasions), with the only difference being the additional subscript to denote individual specific

p, ϕ, and χ terms. The complete data likelihood can then be written as

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Delta}, \tau | \boldsymbol{\Omega}, \boldsymbol{Z}) = \prod_{i=1}^{n} \left(Pr(\boldsymbol{\omega}_i | \mathbf{z}_i, a_i) \times \prod_{t=a_i+1}^{T} Pr(z_{i,t} | z_{i,t-1}) \right).$$

Note that the first term inside the parentheses denotes the probability associated with individual i's capture history while the second term represents the covariate model.

Due to the presence of missing covariate data, this model is typically fit in a Bayesian framework via a Gibbs sampler, an MCMC technique, although maximum likelihood approaches have also been proposed. Two of these methods were briefly introduced in Section 1.1. The first, proposed by Langrock et al. (2013) relies on approximating the integral over all possible missing covariate values by discretizing the range of covariates, similar to the method from Nichols et al. (1992). The key difference, however, is that the range of covariate values is much more finely discretized, and the likelihood is re-formulated as a hidden Markov model to make the estimation much more efficient. The other technique comes from Worthington et al. (2015) and relies on multiple imputation. First, a model is fit to the observed covariates, ignoring any capture history data. Many complete sets of covariates are then generated by sampling missing covariate values from the fitted covariate model. Lastly, the CJS model extended to include continuous covariates is fit separately to each complete set of covariates and the parameter estimates are then aggregated using a non-parametric bootstrap to properly account for uncertainty in the missing covariate values (Buckland, 1984; Buckland and Garthwaite, 1991; Little and Rubin, 2014).

One of the primary reasons these maximum likelihood approaches were developed is that this particular extension of the CJS model can be computationally intensive to fit with MCMC, as missing covariate values must be imputed for every individual at every capture occasion after an individual's first capture on which they are not captured. For extremely large data sets, this can make fitting this model via a Gibbs sampler computationally unfeasible. For example, if I were analyzing a data set in which there were 5,000 instances where individuals were not recaptured after their first capture, fitting this model via a Gibbs sampler would require sampling from 5,000 different full conditional distributions of missing covariates on each iteration of the Gibbs sampler. If I want my Gibbs sampler to generate 3 Markov chains of length 10,000, this would require sampling from the full conditional distributions of the missing covariates 150 million times. Additionally, the full conditional distributions of the missing covariates are rarely in closed form, so the Gibbs sampler will need to be generalized into a Metropolis-Hastings or other rejection sampling algorithm to facilitate sampling from the unknown distributions, making those 150 million sampling procedures even more computationally demanding. The new methods I describe in this manuscript present new ways to reduce this computational burden by first implementing an alternative approach to MCMC, and secondly by defining a new model where fewer missing covariates need to be imputed.

1.5 Cliff Swallows Data

The ability of the two new methods I will present to improve the efficiency of the model fitting algorithms will be evaluated both via simulation studies and the analysis of an actual, large, mark-recapture data set that gives existing methods computational problems. This large mark-recapture data set comes from a 35 year study of cliff swallows led by Dr. Charles R. Brown (Brown and Brown, 1996).

To assess the performance of my two new methods, I will analyze $T = 29$ years of data (each year acting as a capture occasion) collected from 1984 to 2012. A total of 223,092 unique birds were marked during this period. However, we wish to incorporate a weight covariate into our analysis, so captures that did not have a weight covariate associated with them were ignored. Note that if the individual covariates missing on capture occasions where the associated individual was captured are not missing at random, this could produce biased parameter estimates. In addition, to simplify the modeling process, only birds that were banded and observed as adults were included in the analysis. If adolescent birds were included in the analysis, we would likely need to modify the covariate model and treat survival as age category dependent. Removing adolescent birds and captures without associated weight covariates brings the total number of birds in the data set down to $n = 164,621$.

Modeling the effect of an individual, time-varying covariate (weight, in this case) on capture and/or survival probabilities using the approach outlined in Bonner and Schwarz (2006) requires an MCMC algorithm to generate samples from the posterior distribution. Unfortunately, fitting this model to the cliff swallows data set requires the imputation of $1,968,151$ missing covariates on each iteration of the algorithm. Gibbs sampling software packages such as JAGS (Plummer, 2003) or BUGS (Lunn et al., 2000) will therefore require substantial amounts of time and computational resources to generate samples that have converged to the posterior distribution. Later in this manuscript, I fit the CJS model with weight included as a covariate to a small subset of this data ($27,973$ individuals with $56,742$ missing covariate values) via JAGS, and the MCMC algorithm required 18.6 hours to reach convergence. Assuming that the run time scales linearly with the number of missing covariates, I would estimate that fitting the same model to the full cliff swallows data set would require nearly four weeks to reach convergence. Note that this estimate is quite conservative, as the additional missing covariates and capture occasions present in the complete data set would likely result in the algorithm requiring more iterations than the smaller subset to reach convergence. My goal is to reduce this computational burden by first using an alternative to MCMC algorithms that will generate estimates of the posterior distribution more quickly. Then, I will modify the likelihood so that I do not need to impute so many covariate values on each iteration of an MCMC algorithm.

Chapter 2 Variational Bayes

2.1 Introduction to Variational Bayesian Methods

Variational Bayesian methods are an alternative to MCMC for approximating posterior distributions and are often faster while sacrificing some accuracy. Variational Bayesian methods are widespread in computer science, particularly in machine learning applications (Jordan et al., 1999; Jaakkola and Jordan, 2000; Minka, 2001; Mandt and Blei, 2014; Polatkan et al., 2015). However, this approach is starting to gain traction in the statistics literature (Ormerod and Wand, 2010; Kucukelbir et al., 2015). This approach is especially helpful when large data sets render MCMC impractical, as is the situation with the CJS model extended to include individual, time-varying covariates.

Given data ($\mathbf{y}$), parameters ($\boldsymbol{\theta}$), a model ($p(\mathbf{y}|\boldsymbol{\theta})$), and prior distributions on the parameters ($p(\boldsymbol{\theta})$), the goal of variational Bayesian inference is to approximate the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$, with a distribution, $q(\boldsymbol{\theta})$, coming from some restricted class of distributions. The optimal variational distribution $q^*(\boldsymbol{\theta})$ is the member of the restricted class that minimizes the Kullback-Leibler distance between itself and the true posterior. The restricted class of distributions should be chosen such that finding the optimal variational distribution is tractable and the restrictions placed on the variational distributions do not depart too radically from properties of the true posterior distribution.

The critical step in finding the optimal variational distribution is the minimization of the Kullback-Leibler (K-L) distance between $p(\boldsymbol{\theta}|\mathbf{y})$ and $q(\boldsymbol{\theta})$. The K-L distance is defined as:

$$\begin{aligned} KL(q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\mathbf{y})) &= E_{q(\theta)}\left[\log\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\right)\right] \\ &= \int \log\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\right) q(\boldsymbol{\theta}) d\theta. \end{aligned}$$

Minimizing the K-L distance between the true posterior distribution and members of a restricted class of distributions requires algebraic manipulation such that optimizing the K-L distance for $q(\boldsymbol{\theta})$ will not require knowledge of the true posterior distribution. This is a critical step that nearly all variational Bayesian algorithms depend on. Consider that:

$$
\begin{aligned}
KL(q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\mathbf{y})) &= \int \log\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\right) q(\boldsymbol{\theta}) d\theta \\
&= \int \log\left(\frac{q(\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\right) q(\boldsymbol{\theta}) d\theta + \log(p(\mathbf{y})) \int q(\boldsymbol{\theta}) d\theta \\
&= -E_{q(\theta)}\left[\log\left(\frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right] + \log(p(\mathbf{y})).
\end{aligned}
$$

Note that $\log(p(\mathbf{y}))$ does not depend on $q(\boldsymbol{\theta})$, so minimizing the K-L distance between $q(\boldsymbol{\theta})$ and the true posterior distribution is equivalent to maximizing $E_{q(\theta)}\left[\log\left(\frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right]$, an expression which only depends on the the likelihood and prior distribution of $\boldsymbol{\theta}$. This expression is often denoted by $\mathcal{F}[q]$ and can be interpreted as the lower bound on the marginal likelihood (Ormerod and Wand, 2010). Making the maximization of this quantity tractable is the primary concern when selecting a class of variational distributions.

Ormerod and Wand (2010) and McGrory et al. (2009) described several advantages of the variational Bayes methodology relative to MCMC. These advantages include speed (particularly with regards to large data sets), results that are deterministic, and approximate posterior distributions in closed form. Disadvantages of these methods include the fact that they rely on distributional assumptions placed on the variational distribution, $q(\boldsymbol{\theta})$, to make the minimization of the K-L distance tractable. These distributional assumptions can be difficult or impossible to check. Additionally, the variances of the optimal variational distribution typically underestimate the true posterior variance, sometimes radically so, depending on the assumptions made when restricting the family of variational distributions (Grimmer, 2010). These methods are also not nearly as general as MCMC approaches, often requiring quite a bit of analytical work, and while the MCMC algorithm will eventually sample from the true posterior distribution if run for enough iterations, variational Bayesian algorithms produce an approximation of the true posterior distribution.

Mean Field Variational Bayes

One of the most common variational approximations is the mean field variational Bayesian method (MFVB) (Ormerod and Wand, 2010). The key assumption of MFVB is that the joint variational density factorizes such that

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i)$$

where $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_M)$ ia a partition of $\boldsymbol{\theta}$. Note that this class of variational distributions can be thought of as nonparametric, as the product factorization is the only assumption made on $q(\boldsymbol{\theta})$.

The primary reason for this method's popularity is that the mean field method's product restriction results in tractable minimization of the Kullback-Leibler distance. A variational calculus result shows that the optimal variational distribution, denoted $q^*(\boldsymbol{\theta})$, that achieves the minimum Kullback-Leibler distance is

$$q^*(\theta_i) \propto \exp(E_{-\theta_i} \log p(\mathbf{y}, \boldsymbol{\theta})), \text{ for each } i \text{ such that } 1 \leq i \leq M \tag{2.1}$$

where $E_{-\theta_i}$ denotes the expectation operator with respect to $q_{-i}(\boldsymbol{\theta}) = \prod_{j \neq i} q_j(\theta_j)$. It follows that $E_{-\theta_i} \log p(\mathbf{y}, \boldsymbol{\theta})$ is a function of expected values of functions of $\boldsymbol{\theta}_j$ $(j \neq i)$ and parameters from the prior distributions. A detailed proof of this result is presented in Appendix A.1. Note that if a parameter or vector of parameter's full conditional distribution is of known form, then that parameter or vector of parameter's optimal variational density will also be of known form and expectations with respect to that parameter can usually be easily evaluated. The resulting optimal densities, q^*, introduce circular dependencies which can be resolved in an iterative coordinate ascent algorithm in which the variational parameters are repeatedly, sequentially updated until convergence is reached (Ormerod and Wand, 2010). Moreover, conditionally conjugate priors will, by definition, lead to full conditional distributions of known form, leading to a variational Bayesian algorithm with nice analytical properties (Winn and Bishop, 2005) and, coincidentally, will also lead to nice analytical properties for a Gibbs sampler (Gelman et al., 2014, pg. 280).

For illustrative purposes, I present an example of applying the mean field variational Bayesian approach to a very simplistic model presented by Ormerod and Wand (2010): fitting a normal distribution to data with constant mean and variance. Suppose that $Y_1, \ldots, Y_n$ are independent normal random variables with common mean μ and precision τ. To ensure that the prior distributions are conditionally conjugate, μ is assigned a normal prior with mean μ_0 and precision τ_0 while τ is assigned a gamma prior with shape parameter α_0 and rate parameter β_0. Setting $q(\mu, \tau) = q(\mu)q(\tau)$ as the product restriction results in closed form optimal variational densities. The optimal variational density of μ derived via result 2.1 is:

$$\begin{aligned} q^*_\mu(\mu) &\propto \exp(E_\tau \left[\log p(\mathbf{y}|\mu, \tau) + \log p(\mu) + \log p(\tau)\right]) \\ &\propto \exp\left(\sum_{i=1}^{n} -\frac{(y_i - \mu)^2 E_\tau[\tau]}{2} - \frac{(\mu - \mu_0)^2 \tau_0}{2}\right). \end{aligned}$$

Completing the square shows that $q^*_\mu(\mu)$ is proportional to the kernel of a normal distribution:

$$q^*_\mu(\mu) \text{ is } \mathcal{N}\left(\frac{nE_\tau[\tau]\bar{y} + \tau_0\mu_0}{nE_\tau[\tau] + \tau_0}, (nE_\tau[\tau] + \tau_0)^{-1}\right).$$

Similar derivations show that $q^*_\tau(\tau)$ is proportional to the kernel of a gamma distribution:

$$q^*_\tau(\tau) \text{ is } \mathcal{G}\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\sum_{i=1}^{n} E_\mu[(y_i - \mu)^2]\right).$$

Note that the definition of these densities are circular because the optimal variational density for μ depends on the expected value of τ which in turn depends on an expected value with respect to μ. This interdependency is resolved with an iterative algorithm. Let $\mu_{q^*(\mu)}$ represent the mean of the normally distributed variational density for μ, $\tau_{q^*(\mu)}$ represent the precision of the normally distributed variational density for μ, and $\beta_{q^*(\tau)}$ represent the rate parameter of the gamma distributed variational density for τ. The shape parameter of the gamma distributed variational density for τ is $\alpha_{q^*(\tau)} = \alpha_0 + \frac{n}{2}$ and does not need to be included in the iterative algorithm as it does not depend on the variational density of μ. The other three variational parameters need to be included in the iterative algorithm, which I begin by initializing $\beta_{q^*(\tau)}$. Note that $\mu_{q^*(\mu)}$ and $\tau_{q^*(\mu)}$ could also have been initialized first. After initialization, the algorithm can proceed:

$$\begin{aligned}
\tau_{q^*(\mu)} &= n\frac{\alpha_0 + \frac{n}{2}}{\beta_{q^*(\tau)}} + \tau_0 \\
\mu_{q^*(\mu)} &= \left[n\left(\frac{\alpha_0 + \frac{n}{2}}{\beta_{q^*(\tau)}}\right)\bar{y} + \tau_0\mu_0\right]\tau^{-1}_{q^*(\mu)} \\
\beta_{q^*(\tau)} &= \beta_0 + \frac{1}{2}\left(\sum_{i=1}^{n}(y_i - \mu_{q^*(\mu)})^2 + \frac{n}{\tau_{q^*(\mu)}}\right).
\end{aligned}$$

Cycling through these steps leads to rapid convergence.

In this example, the resulting variational densities will be identical to the true marginal posterior distributions due to the fact that the MFVB restriction requiring independence between the variational distributions of τ and μ happens to occur in the true posterior distribution. This is rarely the case in more complex models (Wang and Blei, 2013) and in fact the MFVB algorithm may perform quite poorly if parameters assumed to have independent variational distributions are highly correlated in the true posterior distribution, as I will demonstrate in section 3.1. This can be accounted for by partitioning the variational densities so that highly dependent parameters are grouped together and therefore independence between them is not assumed. However, this will often make it more difficult to derive an efficient algorithm, due to the absence of closed form full conditional distributions, and also requires knowledge of which parameters will be correlated prior to fitting the model.

In Section 2.2, I apply the MFVB method to the CJS model extended to include individual, time-varying covariates, presented in Section 1.4, in an attempt to alleviate the computational burden associated with the traditional MCMC implementation of the model.

Note that there are other variational Bayesian approaches that I considered, such as nonparametric variational inference (Gershman et al., 2012) (which uses a Gaussian mixture to model approximate the posterior distribution) and more parametric approaches that involve specifying parametric distributions for the variational densities and finding a tractable way to maximize the K-L distance. Ultimately, however, these alternative methods did not lead to an algorithm as fast, accurate, or analytically convenient as the mean field approach.

2.2 Application of the Mean Field Approach to the CJS Model with Continuous Covariates

Alteration to the CJS Model with Continuous Covariates

Parameter estimates for the traditional CJS model (without covariates) can be estimated analytically, and for that reason, I immediately apply the variational Bayesian approach to the CJS model extended to allow for continuous, time-varying individual covariates, as described in Section 1.4. When continuous covariates are incorporated into the model, there are no closed form analytical solutions to parameter estimation and Bayesian approaches are implemented (Bonner and Schwarz, 2006). Unfortunately, MCMC algorithms can be computationally unfeasible for large samples and/or many capture occasions. In this section I develop a variational Bayesian algorithm to address this issue.

I use the notation described in Section 1.4 to represent the data and parameters associated with the CJS model extended to include continuous covariates. However, I define a complete data likelihood, adding a latent variable, to make deriving the variational Bayesian algorithm more tractable. Recall that in Section 1.4, $\chi_{i,t}$ represented the probability that individual i was not observed after occasion t (given that the individual was alive on occasion t). Here, rather than summing over all possibilities after an individual's last capture with χ, I consider the time of an individual's death. Let d_i represent the sampling occasion on which individual i was last alive. I can now define a complete data likelihood that does not include any χ terms, but instead relies on the unobserved latent variable d_i to encapsulate the information about an individual after it is last observed. To simplify it's definition, I can factor the complete data likelihood contribution from each individual into two separate components: the first modeling the capture process and the second modeling survival. The capture component of the likelihood for a single individual, i, is defined as:

$$[\boldsymbol{\omega}_i | d_i, \mathbf{p}, f_i] \propto \prod_{t=f_i+1}^{T} (p_t \times \mathbf{1}_{[t \le d_i]})^{\omega_{i,t}} (1 - p_t \times \mathbf{1}_{[t \le d_i]})^{1-\omega_{i,t}}$$

where f_i denotes the capture occasion on which individual i was first captured. The indicator function $\mathbf{1}_{[t \le d_i]}$ is 1 if the individual is alive at time t and 0 otherwise. I then model d_i as a categorical random variable with support 1 through T:

$$[d_i | \boldsymbol{\phi}_i] \sim \text{Categorical}\bigg(1 - \phi_{i,1}, \phi_{i,1}(1 - \phi_{i,2}), ..., (1 - \phi_{i,T-1}) \prod_{t=1}^{T-2} \phi_{i,t}, \prod_{t=1}^{T-1} \phi_{i,t}\bigg)$$

where each term in the above distribution, separated by commas, is a cell probability representing the probability that an individual dies on a particular capture occasion.

The relationship between continuous covariates, such as weight or length, and survival is often of primary interest to researchers. To derive an algorithm that is easy to follow, I focus on models including a single covariate associated with survival and allow capture probabilities to vary across capture occasions. The algorithm below could be extended to include more covariates and covariates associated with capture probabilities relatively easily following the derivation of the MFVB algorithm below as a guide. I employ a logit link function to model the effect of a single, individual, time-varying covariate on survival:

$$\text{logit}(\phi_{i,t}) = \beta_0 + \beta_1 z_{i,t}.$$

Additionally, the missing covariates are modeled as in Bonner and Schwarz (2006), discussed in more detail in Section 1.4:

$$[z_{i,t} | z_{i,t-1}, \Delta_t, \tau] \sim \mathcal{N}\bigg(z_{i,t-1} + \Delta_t, \frac{1}{\tau}\bigg).$$

To complete the specification of the posterior distribution, I assign β_0 and β_1 independent normal priors with means μ_{β_0} and μ_{β_1} and precisions τ_{β_0} and τ_{β_1}, each Δ_t independent normal priors with mean μ_Δ and precision τ_Δ, each p_t independent beta priors with shape parameters α_p and β_p, and τ a gamma prior with shape parameter α_τ and rate parameter β_τ. The hyperparameters that define the prior distributions can be chosen such that the priors are weakly informative, which is the approach I take in the simulations and analysis to follow.

Mean Field Variational Bayesian Algorithm

I define the class of variational distributions by assuming that the posterior distribution of each parameter is independent of the posterior distributions of the other parameters, following the mean field variational Bayes approach introduced in Section 2.1. The only exception is that we model $q(\beta_0, \beta_1)$, allowing the two β parameters to be correlated. This restriction to the class of variational distributions, combined with the new CJS likelihood presented in Section 2.2, yields a product restriction that can be written mathematically as:

$$q(\beta_0, \beta_1, \mathbf{p}, \mathbf{d}, \mathbf{\Delta}, \tau, \mathbf{Z}) = q(\beta_0, \beta_1) q(\tau) \left(\prod_{t=1}^{T} q(p_t) q(\Delta_t) \right) \left(\prod_{i=1}^{n} q(d_i) \right) \left(\prod_{i,t | z_{i,t} \in \mathcal{Z}^{mis}} q(z_{i,t}) \right)$$

where $\mathcal{Z}^{mis}$ denotes the set of all missing covariates (i.e. only the missing covariates have variational densities). Recall that in a Bayesian framework, latent variables and missing data have posterior distributions along with the parameters in the model.

The optimal variational distributions of $\mathbf{d}$, $\mathbf{p}$, $\mathbf{\Delta}$, and τ derived by applying Equation 2.1 are:

1. Variational Distribution of d_i

Categorical with probabilities:

$$q^*(d_i) \propto \begin{cases} \exp(\mathrm{E}_{\beta_0,\beta_1,z_{i,l_i}}[\log(1-\phi_{i,l_i})]), & \text{if } d_i = l_i. \\ \\ \exp(\sum_{t=l_i}^{d_i-1} \mathrm{E}_{\beta_0,\beta_1,z_{i,t}}[\log(\phi_{i,t})] \\ \qquad + \sum_{t=l_i+1}^{d_i} \mathrm{E}_p[\log(1-p_t)] \\ \qquad + \mathrm{E}_{\beta_0,\beta_1,z_{i,d_i}}[\log(1-\phi_{i,d_i})]), & \text{if } d_i = l_i+1, ..., T-1. \\ \\ \exp(\sum_{t=l_i}^{T-1} \mathrm{E}_{\beta_0,\beta_1,z_{i,t}}[\log(\phi_{i,t})] \\ \qquad + \sum_{t=l_i+1}^{T} \mathrm{E}_p[\log(1-p_t)]), & \text{if } d_i = T. \\ \\ 0, & \text{otherwise.} \end{cases}$$

2. Variational Distribution of p_t

Beta with parameters:

$$\alpha_{q^*(p_t)} = \alpha_p + \left(\sum_{i=1}^{N} x_{i,t} \mathbf{1}_{[f_i<t]} \right)$$
$$\beta_{q^*(p_t)} = \beta_p + \left(\sum_{i=1}^{N} (1-x_{i,t}) \mathbf{1}_{[f_i<t]} \mathrm{P}(t \le d_i) \right)$$

3. Variational Distribution for Δ_t

Normal with parameters:

$$\mu_{q^*(\Delta_t)} = \frac{\mathrm{E}_\tau[\tau] \sum_{i|t>f_i} (\mathrm{E}_z[z_{i,t}] - \mathrm{E}_z[z_{i,t-1}]) + \tau_\Delta \mu_\Delta}{\mathrm{E}_\tau[\tau] (\sum_{i|t>f_i} 1) + \tau_\Delta}$$
$$\tau_{q^*(\Delta_t)} = \mathrm{E}_\tau[\tau] \left(\sum_{i|t>f_i} 1 \right) + \tau_\Delta$$

4. Variational Distribution of τ

Gamma with parameters:

$$\alpha_{q^*(\tau)} = \alpha_\tau + \left(\sum_{i=1}^{N} \sum_{t=f_i+1}^{T} \frac{1}{2} \right)$$
$$\beta_{q^*(\tau)} = \beta_\tau + \frac{\sum_{i=1}^{N} \sum_{t=f_i+1}^{T} \mathrm{E}_{\Delta,Z}[(z_{i,t} - z_{i,t-1} - \Delta_t)^2]}{2}$$

The joint variational distribution of β_0 and β_1 is:

$$q^*(\beta_0, \beta_1) \propto \exp\Bigg(\Bigg[\sum_{i=1}^{N} \sum_{t=f_i}^{T} \mathrm{P}(t < d_i)\mathrm{E}_{z_{i,t}}[\log(\phi_{i,t})] + \mathrm{P}(t = d_i)\mathrm{E}_{z_{i,t}}[\log(1 - \phi_{i,t})] \Bigg]$$
$$- \tau_{\beta_0} \frac{(\beta_0 - \mu_{\beta_0})^2}{2} - \tau_{\beta_1} \frac{(\beta_1 - \mu_{\beta_1})^2}{2} \Bigg)$$

This variational distribution does not have a recognizable kernel and therefore must be approximated in some way. To resolve this issue, I have applied a Laplace approximation to construct an approximate multivariate normal density for $q^*(\beta_0, \beta_1)$. Laplace approximations have been implemented previously in variational Bayesian contexts if some of the distributions are not conditionally conjugate (see Wang and Blei (2013)). Assigning $\log q^*(\beta_0, \beta_1)$ as the objective function in a second order Laplace approximation results in an approximate bivariate normal distribution for $q^*(\beta_0, \beta_1)$ with variational parameters

$$\boldsymbol{\mu}_{q^*(\beta_0,\beta_1)} = \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix},$$
$$\boldsymbol{\Sigma}_{q^*(\beta_0,\beta_1)} = - \begin{pmatrix} \frac{\partial^2}{\partial \beta_0^2} \log q^*(\beta_0^*, \beta_1^*) & \frac{\partial^2}{\partial \beta_0 \partial \beta_1} \log q^*(\beta_0^*, \beta_1^*) \\ \frac{\partial^2}{\partial \beta_0 \partial \beta_1} \log q^*(\beta_0^*, \beta_1^*) & \frac{\partial^2}{\partial \beta_1^2} \log q^*(\beta_0^*, \beta_1^*) \end{pmatrix}^{-1}$$

where β_0^* and β_1^* are the values that maximize $\log q^*(\beta_0, \beta_1)$ and

$$\begin{aligned}
\frac{\partial^2}{\partial\beta_0^2} f(\beta_0,\beta_1) &= \Bigg[\sum_{i=1}^{N}\sum_{t=f_i}^{T} -\mathrm{P}(t \le d_i)\mathrm{E}_{z_{i,t}}[\mathrm{expit}(\beta_0+\beta_1 z_{i,t})\times \\
&\qquad (1-\mathrm{expit}(\beta_0+\beta_1 z_{i,t}))]\Bigg] - \tau_{\beta_0} \\
\frac{\partial^2}{\partial\beta_0\partial\beta_1} f(\beta_0,\beta_1) &= \Bigg[\sum_{i=1}^{N}\sum_{t=f_i}^{T} -\mathrm{P}(t \le d_i)\mathrm{E}_{z_{i,t}}[z_{i,t}\mathrm{expit}(\beta_0+\beta_1 z_{i,t})\times \\
&\qquad (1-\mathrm{expit}(\beta_0+\beta_1 z_{i,t}))]\Bigg] \\
\frac{\partial^2}{\partial\beta_1^2} f(\beta_0,\beta_1) &= \Bigg[\sum_{i=1}^{N}\sum_{t=f_i}^{T} -\mathrm{P}(t \le d_i)\mathrm{E}_{z_{i,t}}[z_{i,t}^2\mathrm{expit}(\beta_0+\beta_1 z_{i,t})\times \\
&\qquad (1-\mathrm{expit}(\beta_0+\beta_1 z_{i,t}))]\Bigg] - \tau_{\beta_1}
\end{aligned}$$

and $\mathrm{expit}(x) = (1+e^{-x})^{-1}$, the inverse logit or logistic function.

Similar problems also arise in deriving the variation distributions of the missing covariates. The variational distribution of any missing $z_{i,T}$ (i.e. an individual's covariate value on the last capture occasion of the study) is a normal distribution with variational parameters

$$\begin{aligned}
\mu_{q^*(z_{i,T})} &= \mathrm{E}_z[z_{i,T-1}] + \mathrm{E}_\Delta[\Delta_T] \\
\tau_{q^*(z_{i,T})} &= \mathrm{E}_\tau[\tau]
\end{aligned}$$

Unfortunately, the variational distribution of a missing covariate, $z_{i,t}$, with $t < T$ has an unrecognizable kernel:

$$\begin{aligned}
q^*(z_{i,t}) \propto \exp\Bigg(&- \mathrm{E}_\tau[\tau]\big((z_{i,t}^{mis})^2 + z_{i,t}^{mis}(\mathrm{E}_\Delta[\Delta_{t+1}] - \mathrm{E}_z[z_{i,t+1}] - \mathrm{E}_z[z_{i,t-1}] - \mathrm{E}_\Delta[\Delta_t])\big) \\
&+ \mathrm{P}(t < d_i)\mathrm{E}_{\beta_0,\beta_1}[\log(\mathrm{expit}(\beta_0+\beta_1 z_{i,t}))] \\
&+ \mathrm{P}(t = d_i)\mathrm{E}_{\beta_0,\beta_1}[\log(1-\mathrm{expit}(\beta_0+\beta_1 z_{i,t}))]\Bigg), t < T.
\end{aligned}$$

I again utilize a Laplace approximation, this time separately on each missing $z_{i,t}$ with $t < T$. The resulting approximate variational distribution for each missing covariate, $z_{i,t}$ where $t < T$, then follows a normal distribution with variational parameters

$$\begin{aligned}
&\mu_{q^*(z_{i,t})} = z_{i,t}^* \\
&\sigma^2_{q^*(z_{i,t})} = \big(2\mathrm{E}_\tau[\tau] + \mathrm{P}(t \le d_i)\mathrm{E}_{\beta_0,\beta_1}[\beta_1^2\phi_{i,t}(1-\phi_{i,t})]\big)^{-1}
\end{aligned}$$

where $z_{i,t}^*$ is the value that maximizes $\log(q^*(z_{i,t}))$.

The optimal variational densities derived above have circular dependencies with one another manifested in the expected value terms. Most of these expected values are straightforward to compute. However, there are a few that do not have closed form solutions. In particular, the expected value of any function of $\phi_{i,t} = (e^{\beta_0+\beta_1 z_{i,t}})/(1+e^{\beta_0+\beta_1 z_{i,t}})$ with respect to β_0, β_1, and/or a missing covariate cannot be computed analytically. I approximate these expected values with 1st order Taylor series expansions. As with the simple normal model presented in Section 2.1, the circular dependencies can then be resolved iteratively via Algorithm 1. The expected values necessary to compute the updates in Algorithm 1 can be found in Appendix A.2.

Algorithm 1 Variational Bayesian Algorithm for the Analyses of CJS Models with Individual, Continuous, Time-Varying Covariates

```
Initialize the variational parameters of $q^*(p_j)$, $q^*(\beta_0, \beta_1)$, $q^*(\tau)$, $q^*(\Delta_j)$, and $q^*(z_{i,j})$ for any missing covariates.
while Change in the variational parameters' joint Euclidean norm is greater than the tolerance do
    for $i$ in $1 : n$ do
        Update cell probabilities for $q^*(d_i)$.
    for $j$ in $2 : T$ do
        Update parameters of $q^*(p_j)$.
    Update Laplace approximation for $q^*(\beta_0, \beta_1)$.
        Use numerical optimization to find the mean vector of the Laplace approximation.
        Use the mean vector to compute the variance-covariance matrix.
    for $i$ in $1 : n$ and do
        for $j$ in $f_i : T$ do
            if $z_{i,j}$ is missing then
                Update the Laplace approximation for $q^*(z_{i,j})$.
                    Use numerical optimization to find the mean of the Laplace approximation.
                    Use the mean to compute the variance.
    Update the parameters of $q^*(\tau)$.
    for $j$ in $2 : T$ do
        Update the parameters of $q^*(\Delta_j)$.
```

In practice I found that the missing covariates, updated in steps 10 through 15 of Algorithm 1, converged more slowly than the other parameters in the model due to the fact that each missing covariate's variational density depends on both its neighbors. More precisely, the variational density of a missing covariate, $z_{i,t}$, depends on both $\mathrm{E}_z[z_{i,t-1}]$ and $\mathrm{E}_z[z_{i,t+1}]$ when $t < T$, which are functions of the variational densities of other missing covariates. Because the variational parameters of the missing covariates are updated sequentially, missing covariate updates can potentially include parameter values from the current iteration and the previous iteration. Repeating lines 10 through 15 until the the variational parameters associated with all missing covariates converge is a technique that has proven successful in similar situations (McGrory et al., 2009) and leads to fewer total iterations of the algorithm being required to reach convergence.

Simulation Study

The two most important aspects to examine when assessing the performance of the variational Bayes algorithm are the speed of convergence and the degree of accuracy with which the optimal variational density approximates the true posterior distribution. Speed of convergence is a key metric because the entire motivation behind the variational Bayesian algorithm is to present a faster alternative to the MCMC approach, which struggles with large mark-recapture data sets that contain continuous, individual, time-varying covariates. The accuracy component is necessary to ensure that the resulting optimal variational density is reasonably close to the true posterior distribution.

To evaluate accuracy and speed of convergence, I generated one hundred data sets with 300 individuals observed over 5 capture occasions under two sets of parameter values, one where the capture probabilities were low (0.4), and one where they were high (0.9). For the low capture scenario, I set the true parameter values as $\beta_0 = -1$, $\beta_1 = 1$, $p_t = 0.4$ for all t, $\Delta_t = 0.8$ for all t, and $\tau = 1$. I generated each individual's initial covariate value from a continuous uniform distribution on $(-0.5, 0.5)$, regardless of when the individual was first captured. This set of true parameter values, in conjunction with the distribution of the initial covariate values, results in expected survival probabilities of 0.27 from f_i to $f_i + 1$, 0.45 from $f_i + 1$ to $f_i + 2$, 0.65 $f_i + 2$ to $f_i + 3$, and 0.80 from $f_i + 3$ to $f_i + 4$ where f_i denotes the capture occasion on which individual i was first captured.

Both the MCMC algorithm and the variational Bayesian algorithm were initialized with the same starting values. The initial value for each Δ_t was generated by computing differences in neighboring covariate observations. Initial values for the missing covariates were generated by interpolating with equal step lengths where possible, and the estimates for Δ_t after the last observed covariate. The initial values for each p_t and both β parameters were generated by fitting a CJS model assuming that the missing covariates interpolated above were observed. The MCMC algorithm was implemented via JAGS version 4.2.0, a statistical software package that performs Gibbs sampling to generate samples from posterior distributions (Plummer, 2003). The posterior samples from JAGS were then processed through R to generate summaries, create graphics, and compare to the variational Bayes approach. The variational Bayes algorithm was programmed with R (R Core Team, 2014).

Due to the fact that the MCMC algorithm is converging to the true posterior distribution while the variational Bayesian algorithm is converging to an approximate posterior distribution, some discussion is necessary to describe how speed of convergence was measured. Since the MFVB algorithm is converging to an approximation of the true posterior distribution, $q^*(\theta)$, the speed of the MFVB method was assessed by recording the number of iterations required to produce variational distributions within a specified threshold of the MFVB's optimal variational distribution, $q^*(\theta)$, rather than the true posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$. I approximated the optimal variational distribution as closely as possible by first running the MFVB algorithm given an extremely strict convergence criterion. Similarly, the MCMC approach converges asymptotically to the true posterior distribution, and so the speed of the MCMC algorithm was judged by looking at the number of iterations required for the posterior samples generated by the MCMC method to be as similar to my sample of the true posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, as the MFVB's estimates were to its approximate posterior target, $q^*(\theta)$. The true posterior distribution cannot be computed directly for the CJS model with continuous covariates, so output from an MCMC chain of length 1,000,000 was used as an approximation.

The parameters I focused on to assess speed of convergence are β_0 and β_1, the intercept and slope of the logit of survival probability, as a practitioner fitting this model would be most concerned with those two parameters. The Kullback-Liebler distance between two multivariate normal distributions is leveraged to measure how close each algorithm's current estimate is to it's target estimate $(\widehat{q^*(\theta)}$ in the case of the variational Bayes and $\widehat{p(\theta|\mathbf{y})}$ in the case of MCMC, where the hat symbols denote that these distributions approximate the targets as described in the previous paragraph). The K-L distance was chosen as the distance measure because it has a closed form when comparing two multivariate normal distributions and it can be interpreted as the information lost when one distribution is used to approximate another. It therefore makes sense to use the K-L distance as a tool to gauge how close an approximate posterior density at a particular iteration of an algorithm is to its algorithm's target density (Burnham and Anderson, 1998, pg. 51). The MCMC chains begin with a burn-in of 1000 iterations when generating estimates, as this gives the chains sufficient time to converge to the posterior distribution.

The boxplots in Figure 2.1 show how many iterations were required from each algorithm for the K-L distance of the approximate distributions of β_0 and β_1 to be within 0.001, 0.0001, and 0.00001 units of each algorithm's respective target distribution. The number of iterations on the y-axis of the figures have been log-transformed to make the figures more visually appealing and readable. The boxplots show that for all three convergence criteria, the MFVB algorithm never required more iterations than the MCMC algorithm to reach convergence and that the difference in convergence speed only increases as the criteria becomes more strict. Additionally, the mean of the iterations necessary to reach convergence are reported in the figure. Even for the most strict convergence criterion (0.00001), the MFVB algorithm did not require more than 200 iterations, on average, while the MCMC algorithm required more than 7,000 iterations, on average, to reach the most liberal convergence criterion considered (0.001).

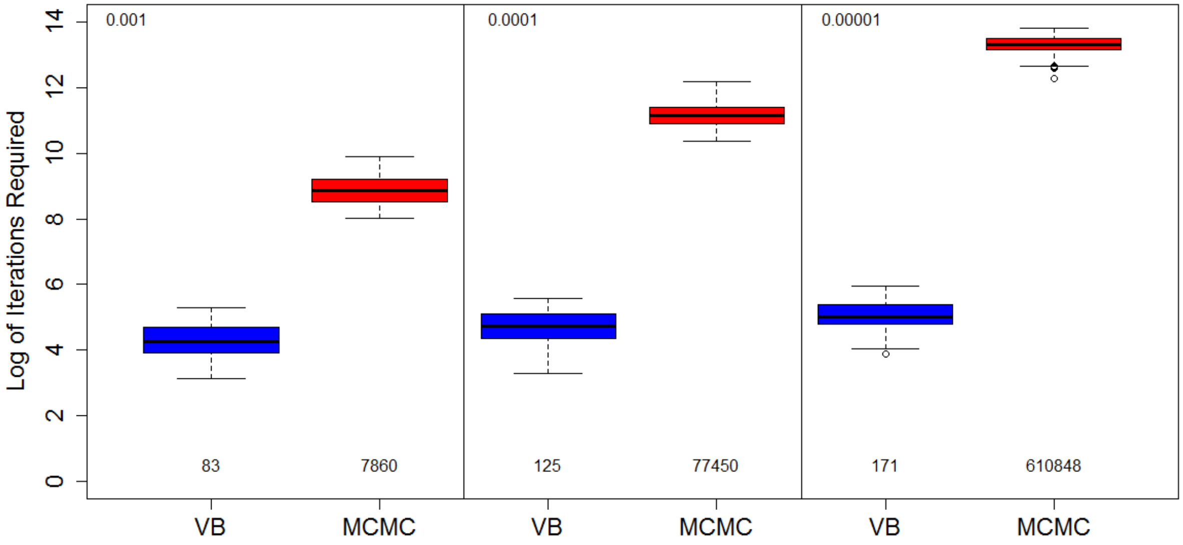


Figure 2.1: *Iterations Required for Convergence: MFVB vs MCMC*

Distribution of the number of iterations required to reach convergence for the MFVB algorithm (blue boxes) and MCMC algorithm (red boxes) in the simulation study. The convergence criterion was set at 10^{-3} (left), 10^{-4} (center), and 10^{-5} (right). The mean number of iterations required to satisfy each convergence criteria is stated under each of the plots.

Now that the efficiency advantage of the MFVB algorithm has been established, the accuracy of the approximate posterior distribution that MFVB algorithm converges to, $q^*(\theta)$, needs to be demonstrated. The coefficients on survival and the capture probabilities for a single simulation replicate are presented in Figure 2.2. The optimal variational distribution generated from the MFVB (in blue) have posterior means that are close to the means of the MCMC generated posterior distribution meant to represent the true posterior. However, the MFVB densities have lower variances than the MCMC densities, consistent with the results from the literature in which posterior distributions approximated via mean field variational Bayesian algorithms typically underestimate the posterior variance (Ormerod and Wand, 2010). This result can be seen even more clearly in Figure 2.3, which displays the target posterior means and standard deviations for both algorithms across all 100 simulated replicates, in addition to the percentage of samples of survival parameters from the true posterior contained in the 95% credible intervals from the optimal variational distribution. Note that this differs from the usual notion of coverage, which concerns the true parameter value. If close to 95% of the posterior samples fall within the 95% credible intervals generated from my variational Bayesian algorithm, this suggests that these credible intervals are similar to those of the true posterior distribution. Similarly, coverage values higher than 95% suggest that my variational Bayesian credible intervals are too wide and coverage values lower than 95% suggest that my variational Bayesian credible intervals are too narrow. The posterior means for the optimal variational distribution and true posterior distribution (approximated by the MCMC algorithm) are very similar while the standard deviations of the target distributions are underestimated by the MFVB results. The underestimation of the true posterior variance is also clear when looking at the proportion of draws from the true posterior distribution (given by the MCMC algorithm) contained in the MFVB generated 95% credible intervals. The proportion of draws contained in the MFVB credible intervals is always under 95% for both β parameters, with most coverage values for β_0 between 0.5 and 0.8 while most of the coverage values for β_1 fall between 0.6 and 0.8, indicating that the credible intervals from the MFVB are much more narrow than those from the true posterior distribution.

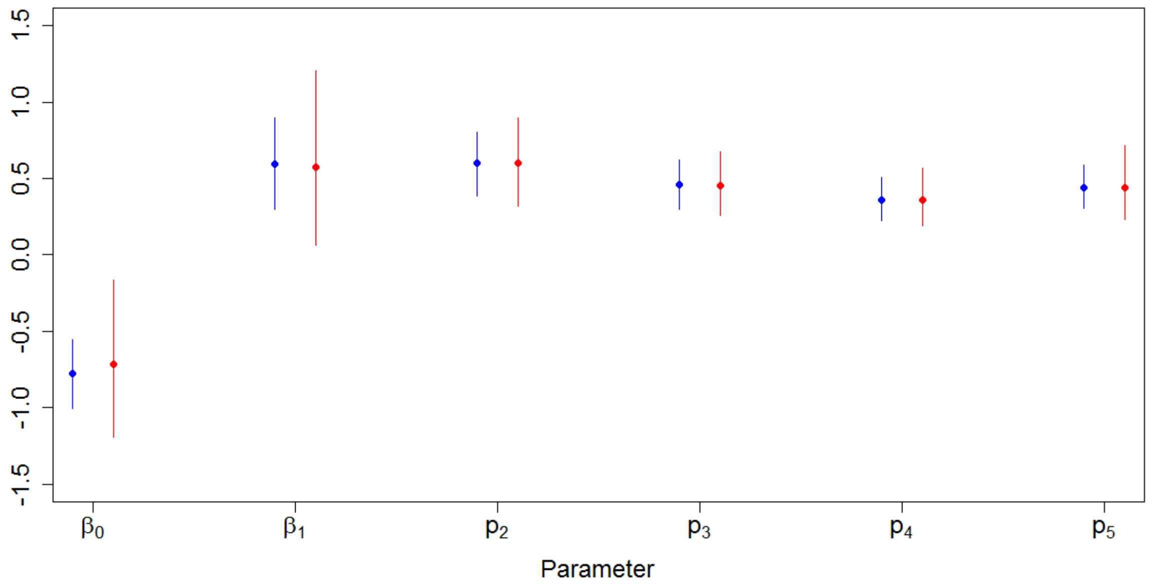


Figure 2.2: *MFVB and MCMC Parameter Estimates in Low Capture Scenario*

MCMC (red) and MFVB (blue) target posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters from a single simulated data set under the low capture scenario ($\boldsymbol{p} = 0.4$). Parameter labels are located on the x-axis.

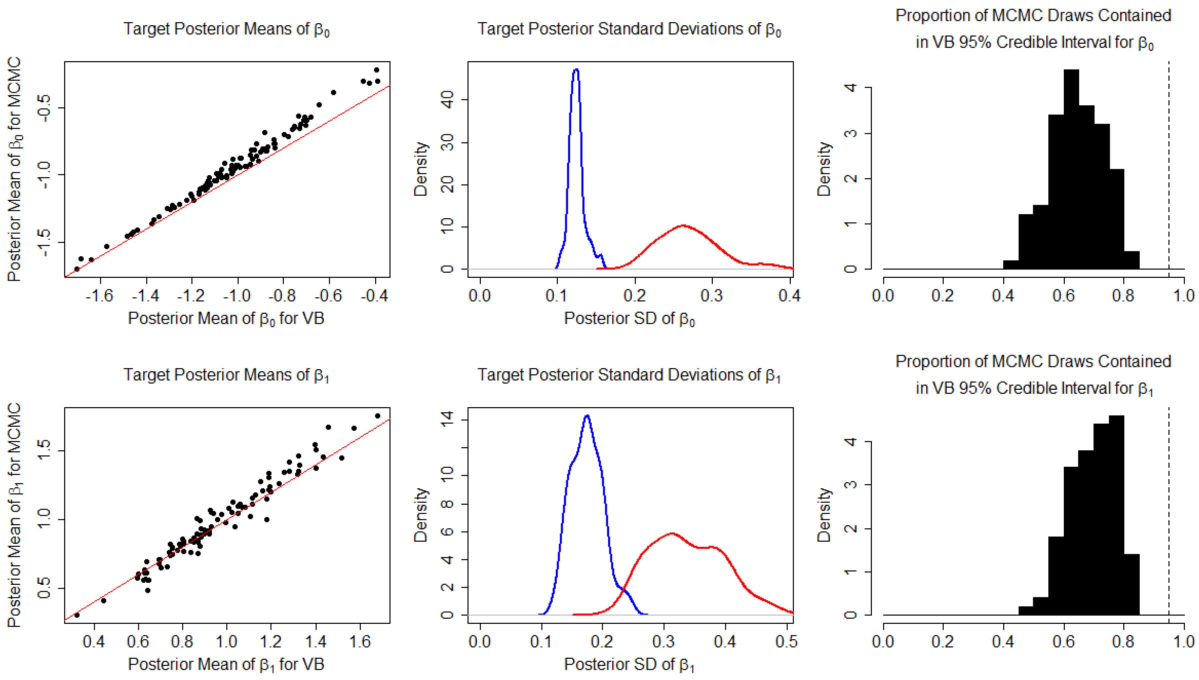


Figure 2.3: *Simulation Results Comparing MFVB and MCMC in Low Capture Scenario*

These plots summarize the 100 simulation results comparing the MFVB and MCMC algorithms under the low capture scenario ($\boldsymbol{p} = 0.4$). The left column of plots shows the relationship between the posterior means of the survival parameters for the MFVB (x-axis) and MCMC (y-axis) algorithms. The center column of plots compares the posterior standard deviations of the capture parameters for the MCMC (red) and MFVB (blue) algorithms. The right column of plots shows histograms of the proportion of MCMC draws that were contained in the 95% credible interval generated by the MFVB algorithm for each simulated data set. Proportions close to 95% would indicate that the posterior distributions generated by the two algorithms match closely.

The underestimation of the posterior variance present when the MFVB method is applied is likely due to the high degree of correlation between the capture and survival parameters in the true posterior distribution. One way to estimate the correlation between parameters in the posterior distribution is to look at the correlation present in the posterior samples generated from a MCMC algorithm. Selecting a random MCMC sample from the aforementioned simulation results shows that β_0 is highly negatively correlated with p_2 ($\rho = -0.44$), p_3 ($\rho = -0.48$), p_4 ($\rho = -0.54$), and p_5 ($\rho = -0.62$) where ρ denotes the Pearson correlation coefficient. This is not terribly surprising, as the product of survival and capture parameters occur quite often in the likelihood function. Unfortunately, this relationship violates a key assumption made when developing my MFVB algorithm: that the survival and capture parameters are independent.

To demonstrate that the correlation between capture and survival parameters is responsible for the underestimation of posterior variance shown in Figures 2.2 and 2.3, I ran a second set of simulations identical to those described previously in this section with one key difference: the capture probabilities were set to 0.9 rather than 0.4. I refer to this as the high capture scenario. Looking at a set of MCMC results from these simulations yields a much smaller set of correlations between β_0 and p_2 ($\rho = -0.12$), p_3 ($\rho = -0.17$), p_4 ($\rho = -0.17$), and p_5 ($\rho = -0.29$). This is because the estimates of the capture and survival parameters become less correlated as the capture probabilities approach 1 due to increased certainty of the reason that an individual is not captured. Additionally, Figures 2.4 and 2.5 show that the posterior approximations generated from the MFVB algorithm are much closer to the true posterior distribution (as approximated by an extremely large MCMC sample). Most of the MFVB 95% credible intervals for the survival parameters now contain between 85% and 94% of the MCMC samples and the posterior means from the optimal variational distribution and true posterior distribution are more similar.

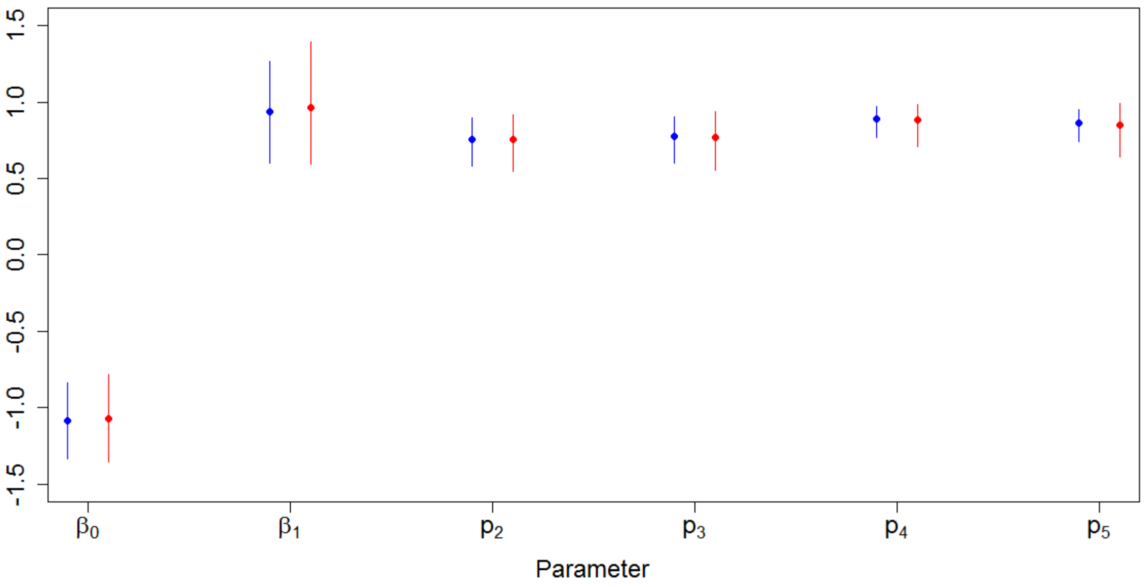
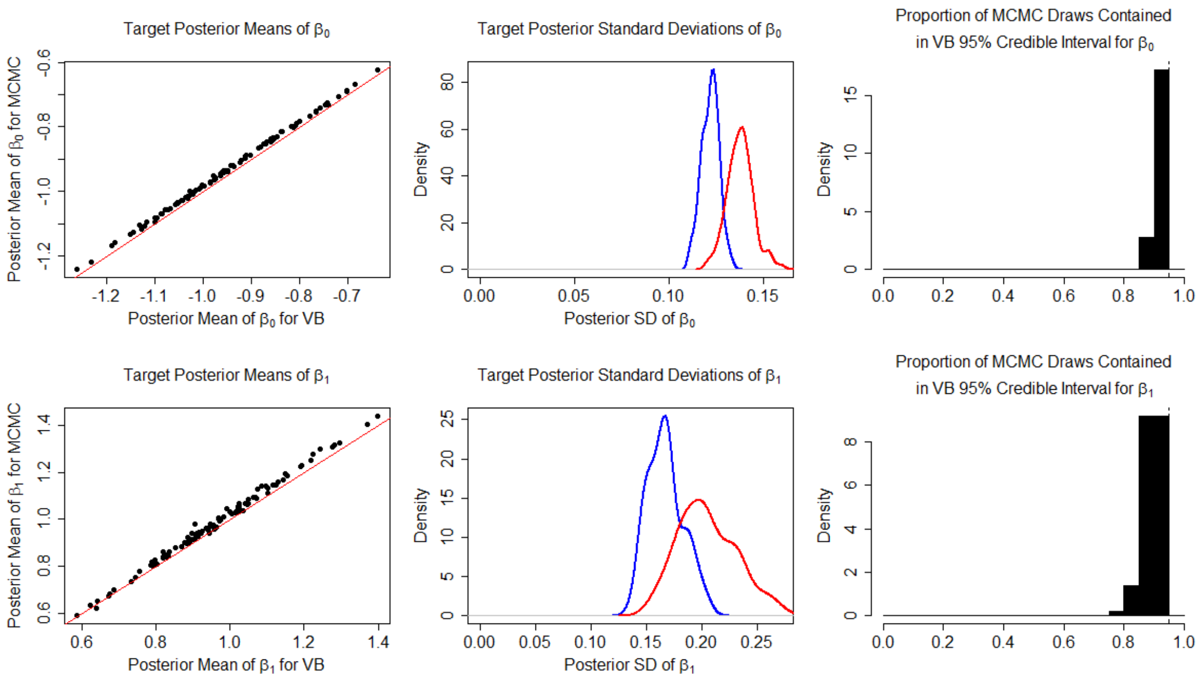


Figure 2.4: *MFVB and MCMC Parameter Estimates in High Capture Scenario*

MCMC (red) and MFVB (blue) target posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters from a single simulated data set under the high capture scenario ($\boldsymbol{p} = 0.9$). Parameter labels are located on the x-axis.


Target Posterior Means of β_0

Target Posterior Standard Deviations of β_0

Proportion of MCMC Draws Contained in VB 95% Credible Interval for β_0

Target Posterior Means of β_1

Target Posterior Standard Deviations of β_1

Proportion of MCMC Draws Contained in VB 95% Credible Interval for β_1

Figure 2.5: *Simulation Results Comparing MFVB and MCMC in High Capture Scenario*

These plots summarize the 100 simulation results comparing the MFVB and MCMC algorithms under the high capture scenario ($\mathbf{p} = 0.9$). The left column of plots shows the relationship between the posterior means of the survival parameters for the MFVB (x-axis) and MCMC (y-axis) algorithms. The center column of plots compares the posterior standard deviations of the capture parameters for the MCMC (red) and MFVB (blue) algorithms. The right column of plots shows histograms of the proportion of MCMC draws that were contained in the 95% credible interval generated by the MFVB algorithm for each simulated data set. Proportions close to 95% would indicate that the posterior distributions generated by the two algorithms match closely.

Discussion

The simulation results indicate that the MFVB algorithm converges much faster than the traditional MCMC approach. However, the approximations are not as accurate as hoped. Examining the posterior samples generated by the MCMC chains, it becomes clear that the posterior distributions of the survival and capture parameters are highly correlated. It seems as though this could be corrected by allowing the variational densities of $\boldsymbol{\beta}$ and $\mathbf{p}$ to be correlated by considering a single variational distribution for $\boldsymbol{\beta}$ and $\mathbf{p}$. The new product restriction would be:

$$q(\boldsymbol{\beta}, \mathbf{p}, \mathbf{d}, \boldsymbol{\Delta}, \tau, \mathbf{Z}) = q(\boldsymbol{\beta}, \mathbf{p})q(\tau)\left(\prod_{t=1}^{T} q(\Delta_t)\right)\left(\prod_{i=1}^{n} q(d_i)\right)\left(\prod_{i,t|z_{i,t}\in\mathcal{Z}^{mis}} q(z_{i,t})\right).$$

Unfortunately, this product restriction does not address the problem, as $\boldsymbol{\beta}$ and $\mathbf{p}$ do not directly appear in the likelihood together. Instead, the correlation arises because they depend on each other through the latent variable d_i, the capture occasion on which individual i was last alive. In other words, this product restriction would still result in independent variational densities for $\boldsymbol{\beta}$ and $\mathbf{p}$ (i.e. $q(\boldsymbol{\beta}, \mathbf{p})$ still factorizes into $q(\boldsymbol{\beta})$ and $q(\mathbf{p})$ using this likelihood and product restriction). Moreover, modeling the joint variational density of $\boldsymbol{\beta}$, $\mathbf{p}$, and $\mathbf{d}$ is not tractable. In the next chapter, I explore potential solutions to this problem by investigating a simpler example where a similar phenomenon occurs.

Chapter 3 Improvements to the Variational Bayesian Algorithm

The correlation of the posterior samples generated by the MCMC algorithm indicates that the assumptions required to implement the MFVB method, namely that the posterior distributions of the capture and survival parameters are independent, cannot accurately approximate the true posterior distribution. To further understand this issue, I investigated the performance of the MFVB method when fitting simpler models featuring posterior distributions with highly correlated parameters that depend on each other via a latent variable (d_i in the case of the CJS model). In particular, I focus on the variational Bayesian implementation of the linear mixed model featured in Ormerod and Wand (2010).

3.1 Mixed Effects Model

Ormerod and Wand (2010) define the general form of a linear mixed model as:

$$\begin{aligned} \mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}, \mathbf{R} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \\ \mathbf{u}|\mathbf{G} &\sim \mathcal{N}(\mathbf{0}, \mathbf{G}) \end{aligned}$$

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{u}$ is a vector of random effects, $\mathbf{X}$ is the design matrix for the fixed effects, $\mathbf{Z}$ is the design matrix for the random effects, $\mathbf{R}$ is the observation level covariance matrix, and $\mathbf{G}$ is the covariance matrix of the random effects. To make the derivations easier to follow, Ormerod and Wand (2010) restrict their model to the variance component model where

$$\mathbf{G} = \text{blockdiag}(\sigma_{u1}^2 \mathbf{I}_{K_1}, \ldots, \sigma_{ur}^2 \mathbf{I}_{K_r})$$

and

$$\mathbf{R} = \sigma_\epsilon^2 \mathbf{I}_n.$$

Additionally, the parameters are given the following priors:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_p) \\ \sigma_{ul}^2 &\sim \text{IG}(A_{ul}, B_{ul}), \quad 1 \leq l \leq r \\ \sigma_\epsilon^2 &\sim \text{IG}(A_\epsilon, B_\epsilon) \end{aligned}$$

where the hyperparameter values can be chosen to make the priors weakly informative.

To implement the mean field variational Bayesian method, Ormerod and Wand (2010) apply a variational density restriction that assumes independence between the variance parameters $(\sigma^2_{u1}, \dots, \sigma^2_{ur}, \sigma^2_\epsilon)$ and the fixed and random effects $(\boldsymbol{\beta}, \mathbf{u})$:

$$q(\boldsymbol{\beta}, \mathbf{u}, \sigma^2_{u1}, \dots, \sigma^2_{ur}, \sigma^2_\epsilon) = q_{\boldsymbol{\beta},\mathbf{u}}(\boldsymbol{\beta}, \mathbf{u}) q_{\boldsymbol{\sigma}^2}(\sigma^2_{u1}, \dots, \sigma^2_{ur}, \sigma^2_\epsilon).$$

Applying equation 2.1 to the linear mixed model with the variational density restriction described above leads to an iterative algorithm with closed form updates. Ormerod and Wand (2010) omit the derivations; however I have provided them in Appendix A.3 for illustrative purposes. The optimal variational distributions are:

$$q^*_{\boldsymbol{\beta},\mathbf{u}} \text{ is } \mathcal{N}(\mu_{q_{\boldsymbol{\beta},\mathbf{u}}}, \Sigma_{q_{\boldsymbol{\beta},\mathbf{u}}})$$

where

$$\mu_{q_{\boldsymbol{\beta},\mathbf{u}}} = \Sigma_{q_{\boldsymbol{\beta},\mathbf{u}}} \left(E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_\epsilon}\right] \mathbf{C}^T \mathbf{y} \right)$$

$$\Sigma_{q_{\boldsymbol{\beta},\mathbf{u}}} = \left(\text{blockdiag}\left(\frac{1}{\sigma^2_{\boldsymbol{\beta}}} \mathbf{I}_p, E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{u1}}\right] \mathbf{I}_{K_1}, \dots, E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{ur}}\right] \mathbf{I}_{K_r} \right) + E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_\epsilon}\right] \mathbf{C}^T \mathbf{C} \right)^{-1}$$

with $\mathbf{C} = [\mathbf{X}, \mathbf{Z}]$, and

$$q^*_{\sigma^2_\epsilon} \text{ is IG} \left(\frac{n}{2} + A_\epsilon, \frac{1}{2}\text{tr}(\mathbf{C}\Sigma_{q(\beta,u)}\mathbf{C}^T) + \frac{1}{2}(\mathbf{y} - \mathbf{C}\boldsymbol{\mu}_{q(\beta,u)})^T(\mathbf{y} - \mathbf{C}\boldsymbol{\mu}_{q(\beta,u)}) + B_\epsilon \right)$$

$$q^*_{\sigma^2_{ul}} \text{ is IG} \left(\frac{K_l}{2} + A_{ul}, \frac{1}{2}\text{tr}(\Sigma_{q(ul)}) + \frac{1}{2}\boldsymbol{\mu}^T_{q(ul)}\boldsymbol{\mu}_{q(ul)} + B_{ul} \right).$$

Note that the optimal variational distributions imply that σ^2_ϵ and each σ^2_{ul} are independent. This restriction was not imposed on the variational distributions, but arises as a result of applying the mean field variational Bayes method. The two different variance parameters only depend on each other through the random effects ($\mathbf{u}$), leading to independent optimal variational densities.

As an example, Ormerod and Wand (2010) apply their variational Bayesian algorithm to analyze the `Orthodont` data set available in the `nlme` R package (Pinheiro et al., 2014), including age and gender as predictors and random intercepts for each subject. Reproducing those results produces Figure 3.1.

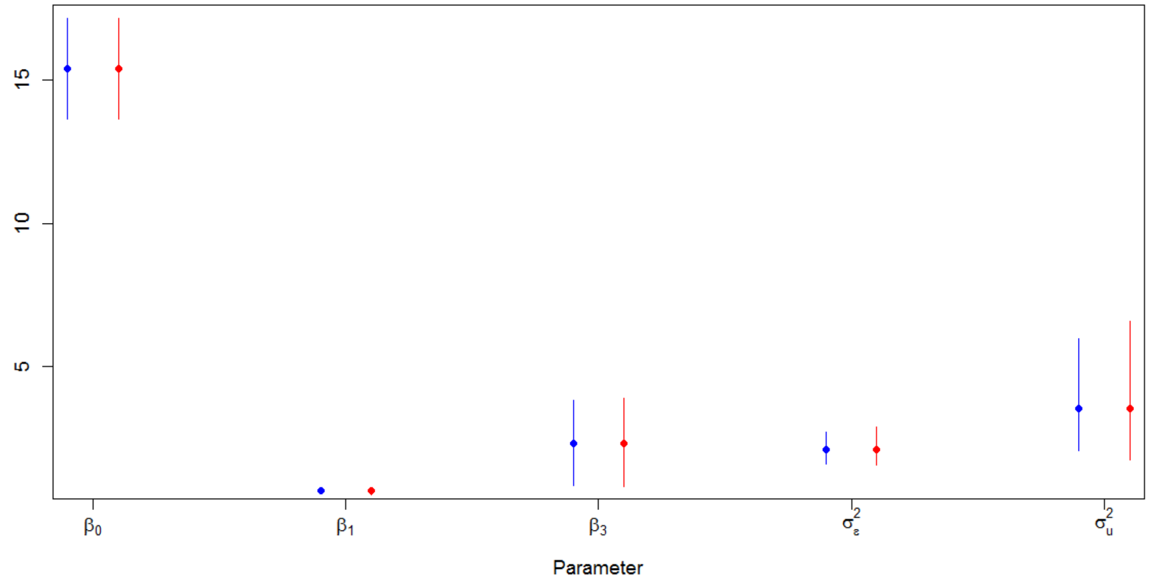


Figure 3.1: *Ormerod and Wand's Algorithm and MCMC Parameter Estimates from Orthodont Data*

Posterior means and 95% credible intervals for Ormerod and Wand's algorithm (blue) and an MCMC implementation (red) of a mixed effects model applied to the Orthodont data set. Parameter labels are located on the x-axis.

Ormerod and Wand's algorithm produces a very good approximation to the true posterior distribution, particularly for the β parameters in this case, and produces these results very quickly (less than one second on my machine). It's possible, however, to generate data where this method does a very poor job of approximating the posterior distributions of the variance parameters. Specifically, this occurs in situations in which the variance parameters are highly correlated *a posteriori*. For example, I generated a single data set from a simple model with no predictors and random intercepts, setting $\sigma^2_\epsilon = 2$ and $\sigma^2_u = 1$ and fitting with an MCMC algorithm. I then observed a posterior correlation of -0.271 between the two variance parameters in my MCMC output (compared to 0.079 for the Orthodont data). Fitting with the MFVB algorithm results in the approximate posteriors displayed in Figure 3.2. The algorithm is severely underestimating the posterior variance of σ^2_u, although the estimate for the intercept is still extremely accurate. This error is caused by the same issues encountered when applying the MFVB method to the CJS model with time varying, individual covariates in Section 2.2. Namely, a high degree of correlation in the posterior distribution between parameters that are uncorrelated conditional on latent data. If I can solve this problem for the MFVB algorithm for the linear mixed effects model, then I can attempt to apply that solution to the more complicated CJS model with time varying, individual covariates.

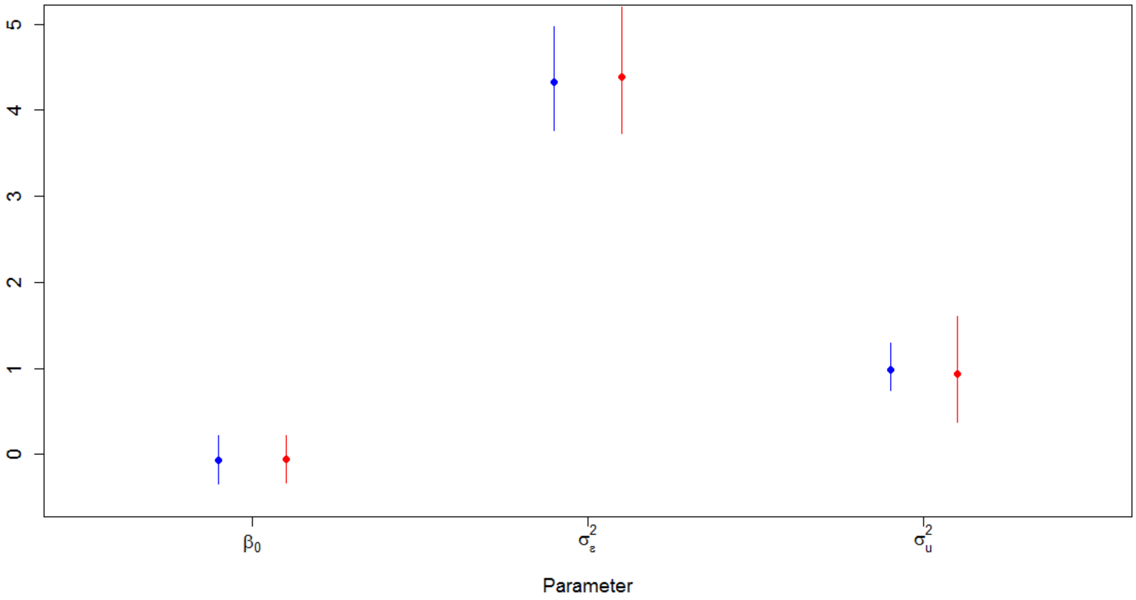


Figure 3.2: *Ormerod and Wand's Algorithm and MCMC Parameter Estimates from Simulated Data*

Posterior means and 95% credible intervals for Ormerod and Wand's algorithm (blue) and an MCMC implementation (red) of a mixed effects model applied to a simulated data set with highly correlated variance parameters. Parameter labels are located on the x-axis.

Allowing Correlation Between Variance Parameters

Recall that the independence of the variational distributions of the variance parameters is not imposed when defining the family of variational distributions, but comes from the fact that the variance parameters only depend on each other through the random effects. This is extremely similar to the problem I encountered when deriving an MFVB algorithm for the CJS model with continuous, time-varying covariates in which my approximate posterior distribution assumes independence between the survival and capture parameters.

In the mixed model situation, correlation between the variance parameters could be incorporated by estimating the joint variational density of both the variance parameters and the random effects, but this makes the derivations too difficult. As an alternative, one can integrate the likelihood over the random effects, which yields a more complicated $\mathbf{R}$ matrix. For example, consider a mixed effects model with only a random intercept. The following formulation of that model is equivalent to Ormerod and Wand's formulation of the mixed model, with the latent random effects vector $\mathbf{u}$ integrated out:

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{R} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{R})$$

where

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{cluster} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{cluster} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \\ \mathbf{0} & \mathbf{0} & & \mathbf{R}_{cluster} \end{pmatrix}$$

and

$$\mathbf{R}_{cluster} = \begin{pmatrix} \sigma_\epsilon^2 + \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\epsilon^2 + \sigma_u^2 & \dots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \\ \sigma_u^2 & \sigma_u^2 & & \sigma_\epsilon^2 + \sigma_u^2 \end{pmatrix}.$$

Additional random effects will result in a more complicated $\mathbf{R}$ matrix that involves covariate values. The non-diagonal $\mathbf{R}$ matrix results in the derivations of the optimal variational densities being more difficult and leads to non-conjugacy, as shown below. Note that the derivations apply to a general covariance matrix $\mathbf{R}$, not necessarily restricted to the random intercept model. The optimal variational density for $\boldsymbol{\sigma^2}$ is:

$$\begin{aligned} q^*_{\boldsymbol{\sigma^2}} &\propto \exp\left(-\frac{1}{2}\log(|\mathbf{R}|) - \frac{1}{2}E_{\boldsymbol{\beta}}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]\right)[\sigma_\epsilon^2][\sigma_u^2] \\ &= \exp\left(-\frac{1}{2}\log(|\mathbf{R}|) - \frac{1}{2}\left(\mathrm{tr}(\mathbf{R}^{-1}\mathbf{X}\Sigma_\beta\mathbf{X}^T) + (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_\beta)^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_\beta)\right)\right)[\sigma_\epsilon^2][\sigma_u^2] \end{aligned}$$

where $[\sigma_\epsilon^2]$ and $[\sigma_u^2]$ represent the prior distributions of σ_ϵ^2 and σ_u^2. When $\mathbf{R}$ is diagonal it is possible to calculate the derminant and the inverse in closed form, isolate the variance parameters, and combine like terms to reveal a product of inverse gamma distributions. The more complex nature of $\mathbf{R}$ in this situation prevents us from applying a similar technique, so I utilize a Laplace approximation to model the variance parameters as a multivariate log-normal distribution. Next, I derive the optimal variational density of the coefficients:

$$q_{\boldsymbol{\beta}}^* \propto \exp\left(E_{\boldsymbol{\sigma}^2}\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]-\frac{1}{2}\beta^T(\sigma_\beta^2\mathbf{I}_p)^{-1}\beta\right).$$

Letting Σ^* denote the element-wise expected value of $\mathbf{R}^{-1}$ with respect to $\boldsymbol{\sigma}^2$, I have

$$q_{\boldsymbol{\beta}}^* \propto \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\Sigma^*(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\frac{1}{2}\beta^T(\sigma_\beta^2\mathbf{I}_p)^{-1}\beta\right).$$

Completing the square yields:

$$q_{\boldsymbol{\beta}}^* \text{ is } \mathcal{N}(\boldsymbol{\mu}_{q_\beta}, \Sigma_{q_\beta})$$

where

$$\Sigma_{q_\beta} = \left((\sigma_\beta^2\mathbf{I}_p)^{-1}+\mathbf{X}^T\Sigma^*\mathbf{X}\right)^{-1}$$

and

$$\boldsymbol{\mu}_{q_\beta} =\Sigma_{q_\beta}\mathbf{X}^{-1}\Sigma^*\mathbf{y}.$$

Note that calculating the element-wise expected value of $\mathbf{R}^{-1}$ is potentially quite difficult. For the random intercept model, deriving the inverse is relatively simple, as it is block-diagonal, and each diagonal has a compound symmetric structure. However, once random slopes or multiple clusters are involved, the $\mathbf{R}$ matrix becomes much more difficult to invert in closed form and calculating the expected value with respect to each element is therefore challenging. For the implementation in the examples to follow, I assume that the expected value of the inverse is the inverse of the expected values, an approach that is not theoretically justified and could potentially lead to issues in approximating the true posterior distribution, but seems to perform well.

Figures 3.3 and 3.4 below compare results from the above method with the method described in Ormerod and Wand (2010) for both the `Orthodont` data set and the simulated data with highly correlated variance parameters. It took longer for the new algorithm, which allows correlation between the variance parameters, to converge (0.4 seconds vs 12.8 seconds for the simple intercept model), due to the numeric optimization routine necessary to compute the Laplace approximation for the variance parameters. However, the new method does estimate the posterior densities of the variance parameters more accurately than the method presented in Ormerod and Wand in the situation in which the variance parameters were highly correlated (see Figure 3.4). The new method also produces results similar to the method presented in Ormerod and Wand when applied to the `Orthodont` data set (see Figure 3.3). The new MFVB algorithm effectively solved the problem by removing the latent variable (the random effects) from the likelihood and allowing the MFVB algorithm to correctly model the correlated parameters. In Section 3.2, I attempt a similar solution by replacing the latent death indicators, d_i, with χ, which effectively sums over all the possible times of death after an individual's last capture.

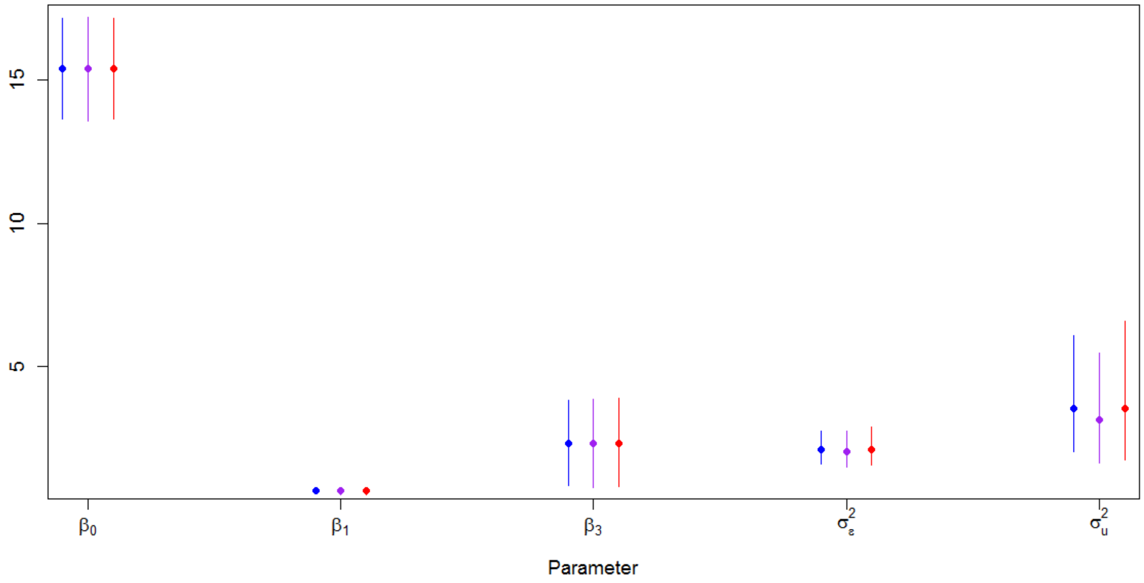


Figure 3.3: *New Algorithm, Ormerod and Wand's Algorithm, and MCMC Parameter Estimates from Orthodont Data*

Posterior means and 95% credible intervals for my new method (purple), Ormerod and Wand's algorithm (blue), and an MCMC implementation (red) of a mixed effects model applied to the Orthodont data set. Parameter labels are located on the x-axis.

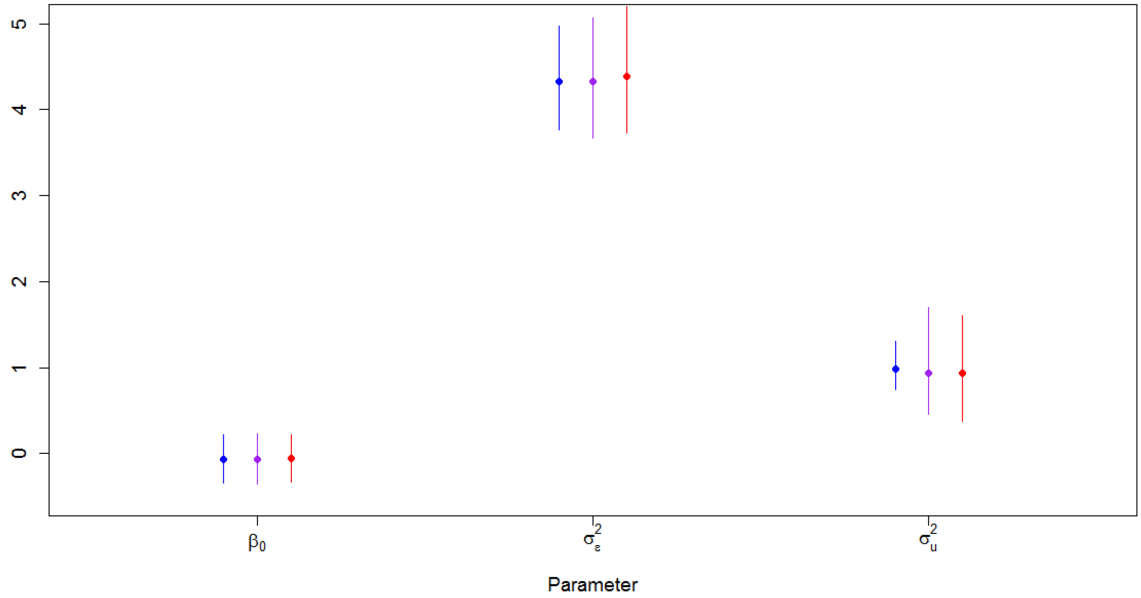


Figure 3.4: *New Algorithm, Ormerod and Wand's Algorithm, and MCMC Parameter Estimates from Simulated Data*

Posterior means and 95% credible intervals for my new method (purple), Ormerod and Wand's algorithm (blue) and an MCMC implementation (red) of a mixed effects model applied to a simulated data set with highly correlated variance parameters. Parameter labels are located on the x-axis.

3.2 Modified Mean Field Approach to the CJS Model with Continuous Covariates

In Section 3.1, I needed to remove a latent variable from the likelihood in order to explicitly model the correlation between two parameters. I follow a similar approach in this section, removing the latent variable representing the time of death of individual i (labeled d_i in Section 2.2). I can allow correlation between $\mathbf{p}$ and $\boldsymbol{\beta}$ in an MFVB algorithm by implementing the product restriction mentioned in Section 2.2 and working with the traditional form of the CJS likelihood defined in Section 1.4 that sums over all possible times of death after an individual's last capture through the χ term. I also model the logit of capture probabilities so that both the capture and survival parameters are on a continuous scale from $-\infty$ to ∞ in order to more straightforwardly approximate the joint variational distribution of $\mathbf{p}$ and $\boldsymbol{\beta}$. The likelihood contribution of individual i is:

$$[\mathbf{x}_i|\mathbf{p}, \boldsymbol{\phi}] \propto \left(\prod_{t=f_i+1}^{l_i} \phi_{i,t-1} p_t^{x_{i,t}} (1-p_t)^{1-x_{i,t}} \right) \chi_{i,l_i}$$

where

$$\begin{aligned} \chi_{i,t} &= (1-\phi_{i,t}) + \phi_{i,t}(1-p_{t+1})\chi_{i,t+1}, \text{ for } t = 1, ..., T-1 \\ \chi_{i,T} &= 1 \\ \text{logit}(p_t) &= \eta_t \\ \text{logit}(\phi_{i,t}) &= \beta_0 + \beta_1 z_{i,t} \end{aligned}$$

and

$$[z_{i,t}|z_{i,t-1}, \Delta_j, \tau] \sim \mathcal{N}\left(z_{i,t-1} + \Delta_t, \frac{1}{\tau} \right).$$

The prior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ do not need to be explicitly defined to derive the algorithm, as the optimal variational densities will need to be approximated and therefore do not require a specific prior distribution to achieve a closed from solution. These prior distributions will be denoted $[\boldsymbol{\beta}]$ and $[\boldsymbol{\eta}]$ in the derivations to follow. In practice, I assign independent normal distributions as priors for β_0, β_1, and each η_t parameter. As in the previous MFVB algorithm for the CJS model with continuous covariates, each Δ_t is assigned an independent, normal prior distribution and τ is assigned a gamma prior distribution.

Recall that in my original MFVB algorithm presented in Section 2.2, I used Laplace's method to approximate the optimal variational distributions of $\boldsymbol{\beta}$ and the missing covariates because they did not have recognizable kernels. Additionally, I used the latent variable d_i (rather than χ) to encapsulate the information about what happens to an individual after their last capture. Unfortunately, when using the product restriction and likelihood defined in this section, the capture parameters can no longer be updated in closed form as they are now incorporated into a joint optimal variational distribution with the survival parameters. This joint optimal variational distribution of capture and survival parameters, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, must be approximated via Laplace's method. The addition of the capture parameters to this variational distribution makes the numerical optimization necessary to estimate the mean more computationally demanding than the earlier algorithm. Additionally, the missing covariates still need to be approximated via a Laplace approximation. However, the addition of the χ term to the likelihood makes this numerical optimization slightly more computationally intensive than before, and, while the difference in computation time is very small for a single missing covariate, these small differences add up to make noticeable difference in the run time of the algorithm (relative to my original MFVB algorithm).

I approximate the optimal variational density of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ as a multivariate normal via a Laplace approximation. First, I follow Equation 2.1 to derive the optimal variational density using the product restriction specified in Section 2.2:

$$\begin{aligned}
q^*_{\eta,\beta} \propto \exp\Bigg(& \sum_{i=1}^{N} \left(\sum_{t=f_i+1}^{l_i} \mathrm{E}_z[\log(\phi_{i,t-1})] + x_{i,t}\log(p_t) + (1-x_{i,t})\log(1-p_t) \right) \\
& + \mathrm{E}_z[\log(\chi_{i,l_i})] + \log([\boldsymbol{\eta}])\log([\boldsymbol{\beta}]) \Bigg) \\
\propto \exp\Bigg(& \sum_{i=1}^{N} \left(\sum_{t=f_i+1}^{l_i} \mathrm{E}_z[\log(\phi_{i,t-1})] + x_{i,t}\log(p_t) + (1-x_{i,t})\log(1-p_t) \right) \\
& + \mathrm{E}_z\left[\log\left((1-\phi_{i,l_i}) + \sum_{j=l_i+1}^{T} \left(\prod_{k=l_i}^{j-1} \phi_{i,k}(1-p_{k+1}) \right)(1-\phi_{i,j}) \right) \right] \\
& + \log([\boldsymbol{\eta}])\log([\boldsymbol{\beta}]) \Bigg).
\end{aligned}$$

Numerical optimization of $\log q^*_{\eta,\beta}$ with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ leads to the mean for the Laplace approximation of the joint variational distribution of $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$. The gradient and Hessian can both be derived in closed form to compute the covariance matrix of the Laplace approximation, but the derivations are unwieldy and have been omitted. Note that this is an approximation of the optimal variational density of $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ under a specific product restriction. The optimal variational density that I am approximating using Laplace's method is itself an approximation of the true posterior distribution, so there are two levels of approximation taking place in this algorithm.

The optimal variational densities of τ and Δ_j remain unchanged from the previous algorithm. However the variational densities of each missing covariate are now different because of changes to the likelihood:

$$\begin{aligned}
q^*_{z^{mis}_{i,t}} \propto \exp\Bigg(& -\mathrm{E}_\tau[\tau]\left((z^{mis}_{i,t})^2 + z^{mis}_{i,t}(\mathrm{E}_\Delta[\Delta_{t+1}] - \mathrm{E}_z[z_{i,t+1}] - \mathrm{E}_z[z_{i,t-1}] - \mathrm{E}_\Delta[\Delta_t]) \right) \\
& + \mathbf{1}_{[t<l_i]}\mathrm{E}_\beta[\log(\mathrm{expit}(\beta_0 + \beta_1 z^{mis}_{i,t}))] \\
& + \mathbf{1}_{[t\geq l_i]}\mathrm{E}_{\beta,z}[\log(\chi_{i,l_i})] \Bigg)
\end{aligned}$$

where $\mathrm{E}_{\beta,z}$ denotes the expected value with respect to $\boldsymbol{\beta}$ and any missing covariate other than $z_{i,t}^{mis}$. Unfortunately, $q^*_{z_{i,t}^{mis}}$ cannot be simplified any more than this and the entire χ_{i,l_i} term must be included when computing the variational distribution of each missing covariate. This distribution is also not in closed form, so a Laplace approximation is applied. The second derivative can be derived so that the covariance matrix of the Laplace approximation can be computed exactly. There is also the issue of the two expected value terms that do not have closed form solutions: $\mathrm{E}_\beta[\log(\mathrm{expit}(\beta_0 + \beta_1 z_{i,t}^{mis}))]$ and $\mathrm{E}_{\beta,z}[\log(\chi_{i,l_i})]$. I approximate these expected values by applying a first order Taylor series expansion. Looking back at Algorithm 1, if I remove lines 3 through 6 that update the optimal variational densities $\boldsymbol{d}$ and $\boldsymbol{p}$ and replace the Laplace approximations of $q^*(\beta_0, \beta_1)$ (line 7) and $q^*(z_{i,j})$ (line 13) with the Laplace approximations of $q^*_{\eta,\beta}$ and $q^*(z_{i,j})$ defined in this section, I have a complete variational Bayesian algorithm that takes into account the correlation between the survival and capture probabilities. This new algorithm, which I call the correlation corrected algorithm, is compared with the original MFVB algorithm I derived in Table 3.1.

I evaluate my new, modified MFVB algorithm in the same manner I evaluated my first MFVB algorithm in Section 2.2. Figure 3.5 shows that, like my original MFVB algorithm, the correlation corrected MFVB algorithm also converges to its target distribution much faster than the MCMC algorithm. For the most strict convergence criterion I considered (0.00001), the correlation corrected MFVB algorithm needed an average of 283 iterations to converge. Meanwhile, the MCMC algorithm, applying the most liberal convergence criterion I considered (0.001), required an average of 7830 iterations to converge to its target distribution (as approximated by an MCMC algorithm run for 1,000,000 iterations).

Table 3.1: *Differences between the original MFVB algorithm and the correlation corrected MFVB algorithm*

Original MFVB Algorithm	**Correlation Corrected MFVB Algorithm**
The survival and capture parameters are assumed to be independent.	The survival and capture parameters are not assumed to be independent.
d is used to encapsulate the information about what happens after an individual's last capture.	χ is used to encapsulate the information about what happens after an individual's last capture.
The capture parameters have optimal variational distributions of known form.	The capture parameters are included in a joint variational distribution with the survival parameters and this joint distribution does not have known form.
The optimal variational distribution of the missing covariates is not of known form.	The optimal variational distribution of the missing covariates is not of known form and the numerical optimization required by the Laplace approximation takes longer due to the addition of the χ term used in the likelihood.

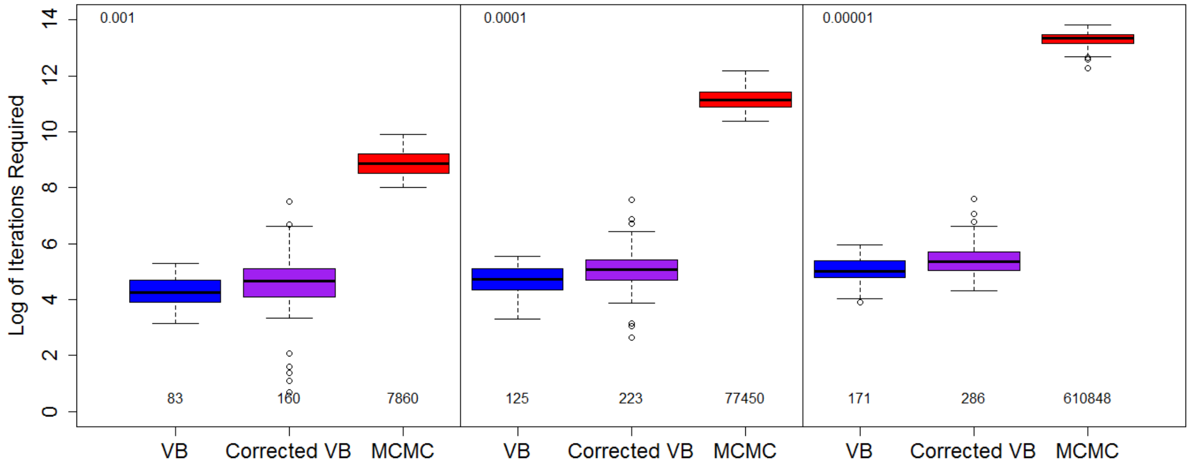


Figure 3.5: *Iterations Required for Convergence: MFVB vs Corrected MFVB vs MCMC*

Distribution of the number of iterations required to reach convergence for the MFVB algorithm (blue boxes), MCMC algorithm (red boxes), and the correlation corrected MFVB algorithm (purple boxes) in the simulation study. The convergence criterion was set at 10^{-3} *(left),* 10^{-4} *(center), and* 10^{-5} *(right). The mean number of iterations required to satisfy each convergence criteria is stated under each of the plots.*

Comparing the number of iterations required to reach convergence of the correlation corrected MFVB algorithm with my original MFVB algorithm referenced in Figure 2.1, it is clear that the correlation corrected MFVB requires more iterations to converge. Furthermore, because of the additional numeric optimization necessary in the correlation corrected MFVB algorithm, the correlation corrected MFVB algorithm will take slightly more time during each iteration than my original MFVB algorithm (0.5 seconds vs. 0.8 seconds per iteration, on average for these simulations). In summary, the correlation corrected MFVB requires fewer iterations to converge to its target distribution than the MCMC algorithm for these simulated data sets, but converges more slowly than my original MFVB algorithm.

The additional time required for the correlation corrected MFVB algorithm to converge to its target distribution does yield some significant advantages with regard to accuracy. Figure 3.6 displays the densities of the capture and survival parameters from the approximate posterior distributions generated from both the MCMC and correlation corrected MFVB algorithms for one randomly selected simulated data set.

Comparing this plot to Figure 2.2 shows that the posterior variability estimated by the correlation corrected MFVB is a much closer match to the true posterior distribution (estimated by the MCMC results) than my original MFVB algorithm. Figure 3.7 shows that the posterior means of the survival parameters are just as accurate as they would be applying the previous algorithm most of the time. However, note that there were a few generated data sets that resulted in posterior means for β_1 that were not very similar to the MCMC results. This is due to the numerical optimization technique necessary in the Laplace approximation reaching local optima and is certainly a drawback to this algorithm, although this does not happen frequently. However, Figure 3.7 does show that the posterior standard deviations are a significantly better match to the MCMC results, and the MCMC draws are contained in the 95% correlation corrected MFVB credible intervals between 85% to 95% of the time. Similar results can be found for the scenario in which $p_2 = p_3 = p_4 = p_5 = 0.9$ in Figures 3.8 and 3.9. However, with capture probabilities of 0.9, I no longer see the numerical optimization issues I experienced when the capture probabilities used to generate the data were set at 0.4.

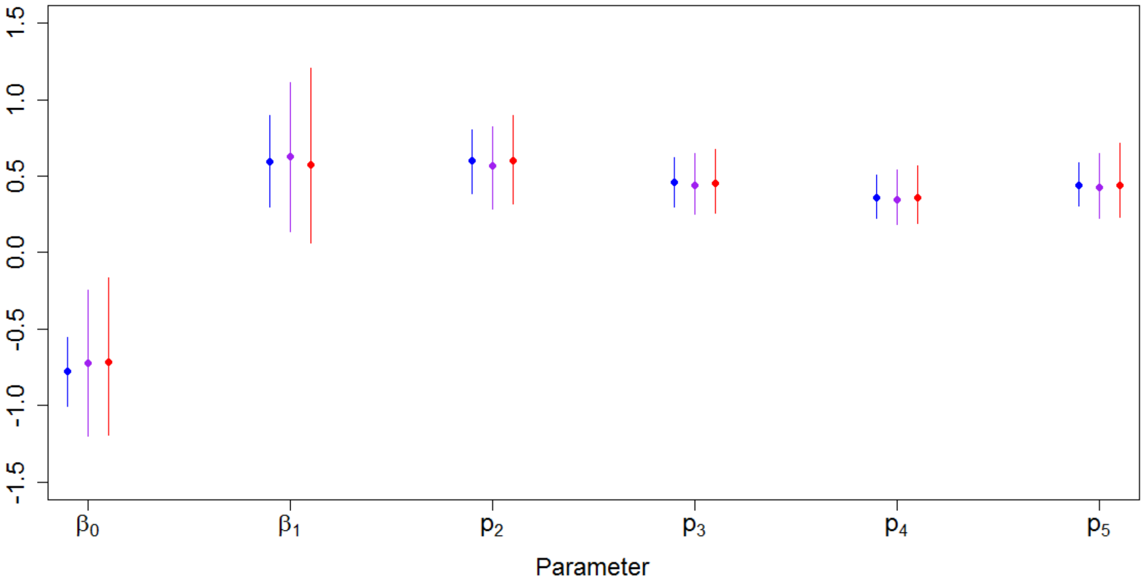


Figure 3.6: *MFVB, Corrected MFVB, and MCMC Parameter Estimates in Low Capture Scenario*

MCMC (red), MFVB (blue), and correlation corrected MFVB (purple) target posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters from a single simulated data set under the low capture scenario ($\boldsymbol{p} = 0.4$). Parameter labels are located on the x-axis.

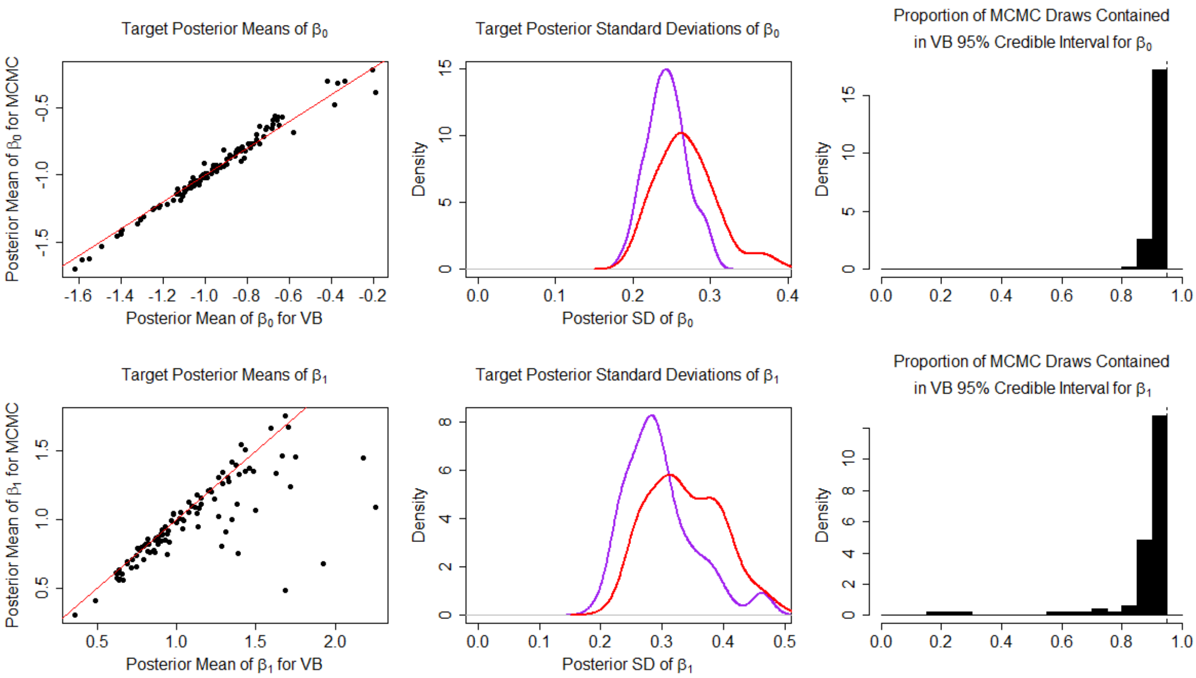


Figure 3.7: *Simulation Results Comparing MFVB, Corrected MFVB, and MCMC in Low Capture Scenario*

These plots summarize the 100 simulation results comparing the correlation corrected MFVB and MCMC algorithms under the low capture scenario ($\mathbf{p} = 0.4$). The left column of plots shows the relationship between the posterior means of the survival parameters for the correlation corrected MFVB (x-axis) and MCMC (y-axis) algorithms. The center column of plots compares the posterior standard deviations of the capture parameters for the MCMC (red) and correlation corrected MFVB (purple) algorithms. The right column of plots shows histograms of the proportion of MCMC draws that were contained in the 95% credible interval generated by the correlation corrected MFVB algorithm for each simulated data set. Proportions close to 95% would indicate that the posterior distributions generated by the two algorithms match closely.

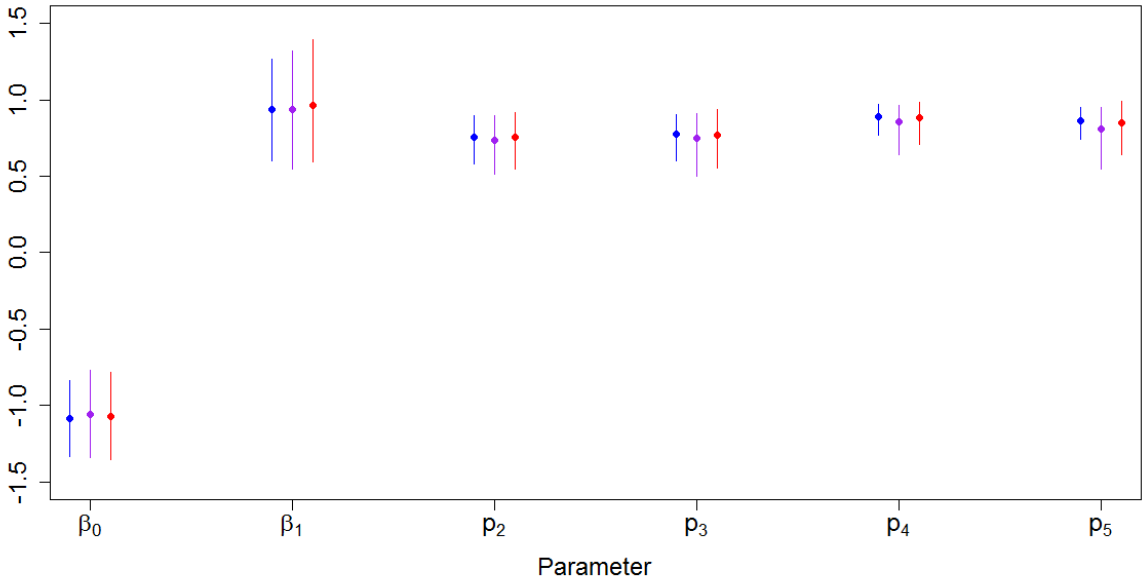


Figure 3.8: *MFVB, Corrected MFVB, and MCMC Parameter Estimates in High Capture Scenario*

MCMC (red), MFVB (blue), and correlation corrected MFVB (purple) target posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters from a single simulated data set under the high capture scenario ($\boldsymbol{p} = 0.9$). Parameter labels are located on the x-axis.

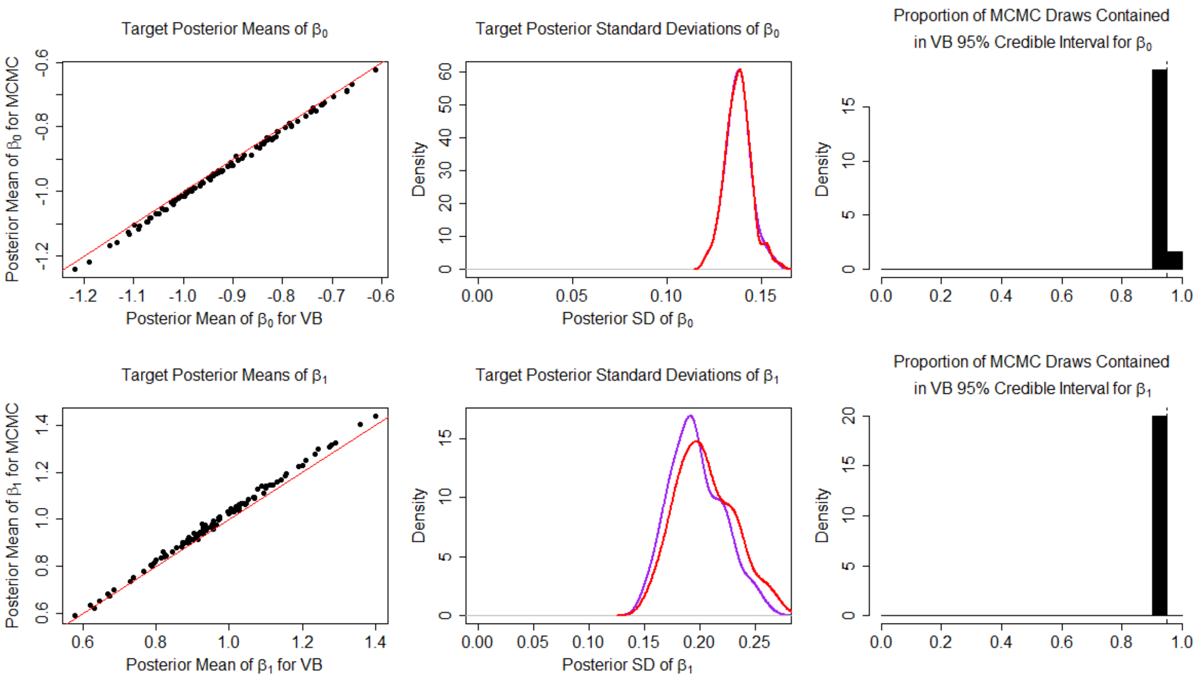


Figure 3.9: *Simulation Results Comparing MFVB, Corrected MFVB, and MCMC in High Capture Scenario*

These plots summarize the 100 simulation results comparing the correlation corrected MFVB and MCMC algorithms under the high capture scenario ($\mathbf{p} = 0.9$). The left column of plots shows the relationship between the posterior means of the survival parameters for the correlation corrected MFVB (x-axis) and MCMC (y-axis) algorithms. The center column of plots compares the posterior standard deviations of the capture parameters for the MCMC (red) and correlation corrected MFVB (purple) algorithms. The right column of plots shows histograms of the proportion of MCMC draws that were contained in the 95% credible interval generated by the correlation corrected MFVB algorithm for each simulated data set. Proportions close to 95% would indicate that the posterior distributions generated by the two algorithms match closely.

3.3 Hybrid Algorithm

Allowing correlation in the variational densities between the capture and survival parameters increases the accuracy of the approximate posterior distributions, matching MCMC results much more closely than my original MFVB algorithm, particularly with regard to the posterior variances. This increased accuracy, however, comes with a cost: significantly longer run time compared to my original MFVB algorithm. There are some situations in which the modified MFVB algorithm can converge even more slowly than the MCMC algorithm, due to the computationally demanding nature of the numerical optimization routine necessary to find the mean vector of the Laplace approximation of $q^*_{\eta,\beta}$. Additionally, the modified algorithm will scale poorly as the number of capture occasions increase, as $\boldsymbol{\eta}$ will grow in dimension and the numerical optimization necessary to update $q^*_{\eta,\beta}$ will be even more computationally intensive. In this section, I attempt to combine the two previously derived algorithms to create a hybrid algorithm which possesses the speed of the original MFVB approach and some of the improved accuracy present in the modified MFVB algorithm.

The idea behind the hybrid algorithm is to run the original MFVB algorithm until convergence, and then to run a final iteration using the modified MFVB algorithm. Looking at the left-most column in Figures 2.3 and 2.5, it is apparent that the original MFVB algorithm estimates the posterior means of the capture parameters quite well. By running the final iteration using the modified MFVB algorithm, the posterior variances can be more accurately estimated without running the computationally intensive updates of $q^*_{\eta,\beta}$ every iteration while waiting for the optimal variational distributions of the missing covariates, $\boldsymbol{\Delta}$, and τ to converge.

Using the same simulated data sets described in detail in Section 2.2 and referred to again in Section 3.2, I can ascertain the effectiveness of this approach. First, note that I do not include boxplots comparing iterations required to reach convergence. This is due to the fact that the hybrid algorithm takes exactly one more iteration that my first MFVB algorithm detailed in Section 2.2. You may refer to Figure 2.1 for a comparison of convergence speed between the hybrid MFVB algorithm and the MCMC algorithm.

A comparison of the approximate posterior distribution computed by the hybrid MFVB algorithm, the results obtained via the MCMC approach, and the previously described MFVB algorithms for a single data set simulated under the the low capture scenario can be viewed in Figure 3.10. The posterior means from the hybrid algorithm are closely aligned with the posterior means generated from the other three algorithms across all six parameters, and the posterior variances (represented by the 95% credible intervals) demonstrate that the hybrid algorithm is much closer to the correlation corrected MFVB and MCMC estimates. Similar results can be seen for a data set simulated under the high capture scenario in Figure 3.11.

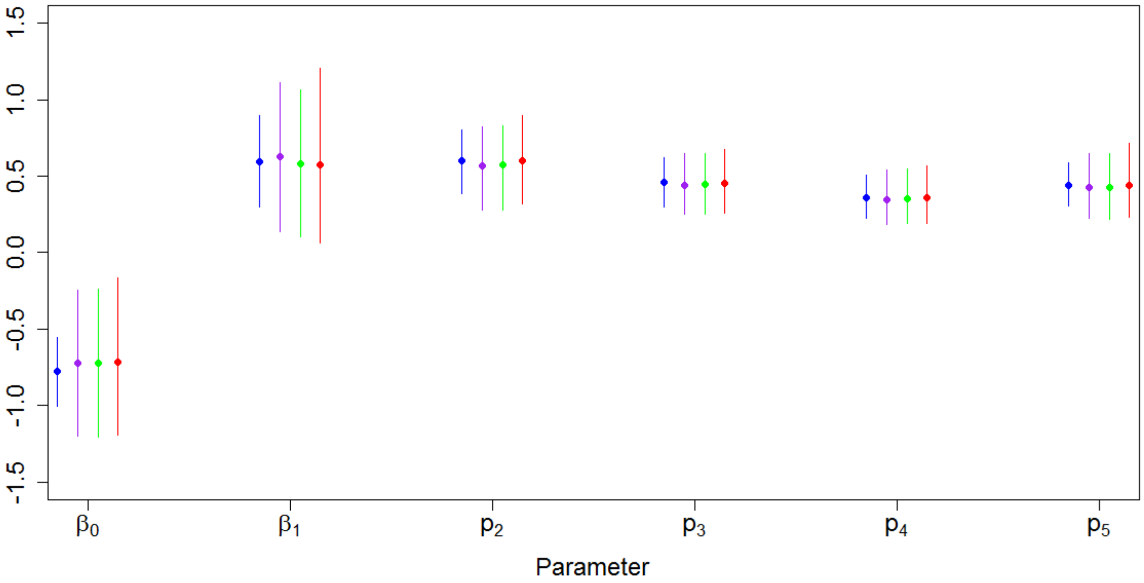


Figure 3.10: *MFVB, Corrected MFVB, Hybrid MFVB, and MCMC Parameter Estimates in Low Capture Scenario*

MCMC (red), MFVB (blue), correlation corrected MFVB (purple), and hybrid MFVB (green) target posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters from a single simulated data set under the low capture scenario ($\boldsymbol{p} = 0.4$)*. Parameter labels are located on the x-axis.*

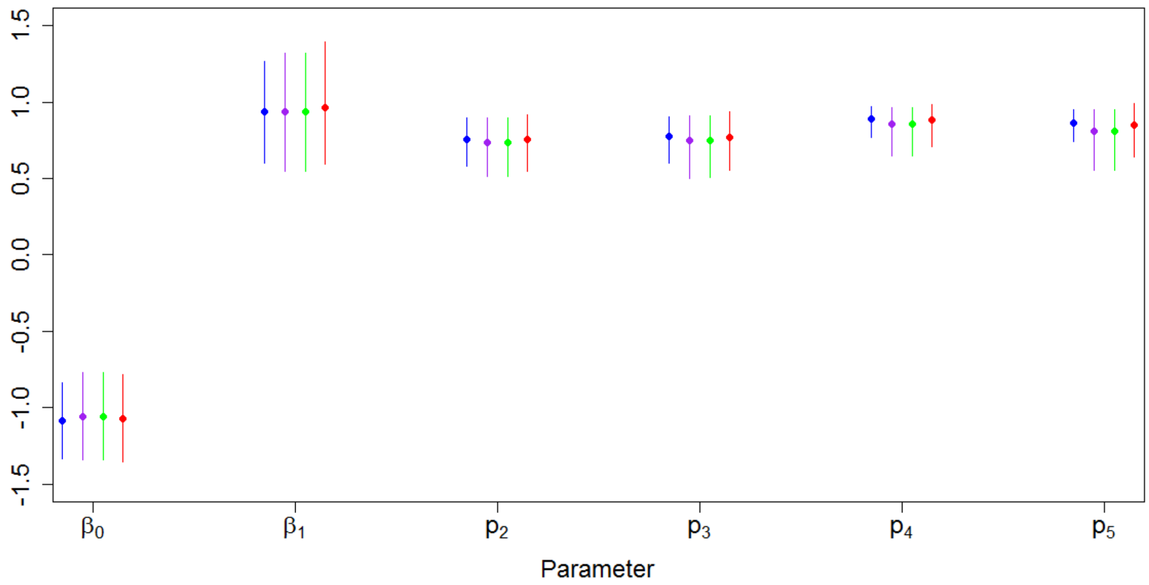


Figure 3.11: *MFVB, Corrected MFVB, Hybrid MFVB, and MCMC Parameter Estimates in High Capture Scenario*

MCMC (red), MFVB (blue), correlation corrected MFVB (purple), and hybrid MFVB (green) target posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters from a single simulated data set under the high capture scenario ($\boldsymbol{p} = 0.9$). Parameter labels are located on the x-axis.

The performance of the hybrid MFVB algorithm across all 100 simulated data sets can be assessed by examining Figures 3.12 (low capture scenario) and 3.13 (high capture scenario). The posterior means of the capture parameters ($\boldsymbol{\beta}$) from the hybrid MFVB algorithm match very closely with the posterior means generated via MCMC. Additionally, the posterior standard deviations from the hybrid MFVB algorithm are more closely aligned with the MCMC results than the output from the original MFVB algorithm were. Also, the proportion of MCMC draws contained in the 95% hybrid MFVB credible intervals are almost all between 0.85 and 0.95 for the low capture scenario and 0.90 and 0.95 for the high capture scenario. These findings are quite similar to the results presented in Figures 3.7 and 3.9 in the correlation corrected MFVB section. However, the hybrid algorithm is significantly faster than the full correlation corrected algorithm (on average, 2.0 minutes vs 6.5 minutes for these simulations) due to the fact that the computationally intensive numerical optimization required to update $q^*(\boldsymbol{\beta}, \boldsymbol{p})$ only occurs on the final iteration. Additionally, in the low capture scenario I did not observe any of the convergence issues that affected the correlation corrected algorithm in Figure 3.7. This can be attributed to the hybrid algorithm's use of the simpler, original MFVB algorithm to converge to reasonable approximate posterior means before utilizing one iteration of the correlation corrected algorithm to more accurately reflect the posterior variance. Relying on high dimensional numerical optimization from the very beginning is what led to the convergence issues with the correlation corrected algorithm, and the hybrid algorithm avoids this problem.

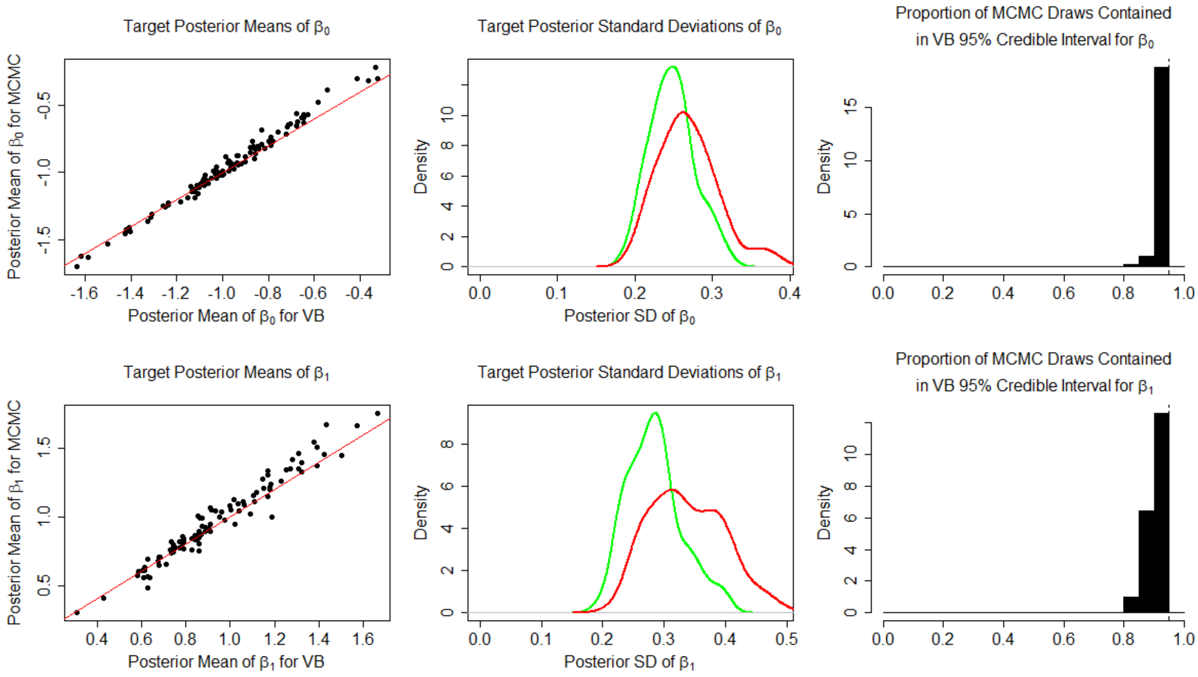


Figure 3.12: *Simulation Results Comparing MFVB, Corrected MFVB, Hybrid MFVB, and MCMC in Low Capture Scenario*

These plots summarize the 100 simulation results comparing the hybrid MFVB and MCMC algorithms under the low capture scenario ($\boldsymbol{p} = 0.4$). The left column of plots shows the relationship between the posterior means of the survival parameters for the hybrid MFVB (x-axis) and MCMC (y-axis) algorithms. The center column of plots compares the posterior standard deviations of the capture parameters for the MCMC (red) and hybrid MFVB (green) algorithms. The right column of plots shows histograms of the proportion of MCMC draws that were contained in the 95% credible interval generated by the hybrid MFVB algorithm for each simulated data set. Proportions close to 95% would indicate that the posterior distributions generated by the two algorithms match closely.

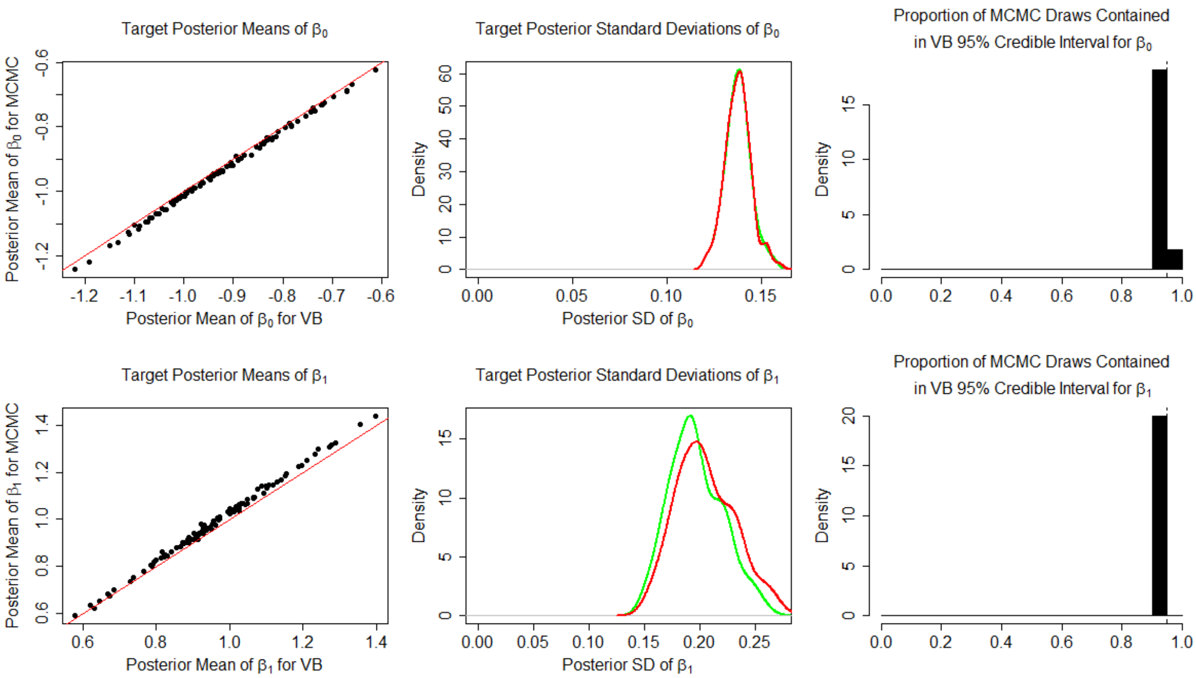


Figure 3.13: *Simulation Results Comparing MFVB, Corrected MFVB, Hybrid MFVB, and MCMC in High Capture Scenario*

These plots summarize the 100 simulation results comparing the hybrid MFVB and MCMC algorithms under the high capture scenario ($\boldsymbol{p} = 0.9$). The left column of plots shows the relationship between the posterior means of the survival parameters for the hybrid MFVB (x-axis) and MCMC (y-axis) algorithms. The center column of plots compares the posterior standard deviations of the capture parameters for the MCMC (red) and hybrid MFVB (green) algorithms. The right column of plots shows histograms of the proportion of MCMC draws that were contained in the 95% credible interval generated by the hybrid MFVB algorithm for each simulated data set. Proportions close to 95% would indicate that the posterior distributions generated by the two algorithms match closely.

The hybrid algorithm yields an approximate posterior distribution extremely similar to the correlation corrected MFVB algorithm, and converges almost 3 times faster (based on my simulation results). Additionally, both of these algorithms generate approximate posterior distributions much closer in posterior variance to the MCMC results than my original MFVB algorithm could. This suggests that the computationally demanding optimization step necessary to update $q^*(\boldsymbol{\beta}, \boldsymbol{p})$ need not occur on every iteration of the algorithm. The missing covariates and the parameters associated with the change in covariates over time ($\boldsymbol{\Delta}$ and τ) can converge to reasonable variational parameter estimates using the more naive updates for the (assumed independent) optimal variational densities of $\boldsymbol{\beta}$ and $\boldsymbol{p}$ present in the original MFVB algorithm. Using the update to $q^*(\boldsymbol{\beta}, \boldsymbol{p})$ that allows the capture and survival parameters to be correlated at the very end of the original MFVB algorithm to correct underestimated posterior variances results in an algorithm that produces both rapid convergence and a more accurate approximate posterior distribution. In essence, this algorithm combines the two most attractive features from the original MFVB algorithm and the correlation corrected MFVB algorithm.

3.4 Application to Cliff Swallows

The simulated data sets I have presented thus far have been relatively small mark-recapture data sets ($n = 300$ and $T = 5$). While this is a reasonable criteria to compare the variational Bayesian results with the more traditional MCMC approach, my purpose in exploring variational Bayesian methods was to analyze extremely large data sets that MCMC algorithms would not be able to handle. In this section, I analyze a very large mark-recapture data set to demonstrate that the MFVB approach is applicable to problems of this magnitude.

Recall the cliff swallows study described in Section 1.5 that consisted of $n = 164,621$ birds observed over $T = 29$ capture occasions. The model I have fit to this data set includes time-varying capture probabilities and an effect of weight on survival, similar to the models examined in Sections 2.2, 3.3 and 3.2. Weights were standardized before analysis. The first 8 years of data were analyzed initially to assess how well the variational Bayesian algorithms performed compared to an MCMC approach. The entire 29 years of data were then analyzed with the original MFVB algorithm and the hybrid algorithm. The MCMC and correlation corrected MFVB algorithm were not used to analyze the entire 29 years worth of data, as this would not be computationally feasible. For the MCMC analysis on the first 8 years of data, I generated 3 chains each of length $50,000$, discarding the first $2,000$ as burn-in. JAGS version 3.4.0 was used to generate the posterior samples (Plummer, 2003), which were then processed through R to create summaries, tables, and graphics (R Core Team, 2014).

The posterior means and 95% credible intervals generated by each of the methods for the first 8 years of cliff swallows data can be viewed in Figure 3.14. Note that all three MFVB methods match the MCMC results closely with regard to the capture probabilities, however the original MFVB algorithm severely underestimates the posterior variance of the survival parameters ($\boldsymbol{\beta}$), as I expected based on my previously presented simulation studies. In contrast, the correlation corrected MFVB and hybrid MFVB methods generate 95% credible intervals very similar in length to the MCMC produced credible intervals. Surprisingly, the hybrid algorithm also appears to produce posterior mean estimates for the survival parameters closer to the MCMC results than those of the correlation corrected MFVB.

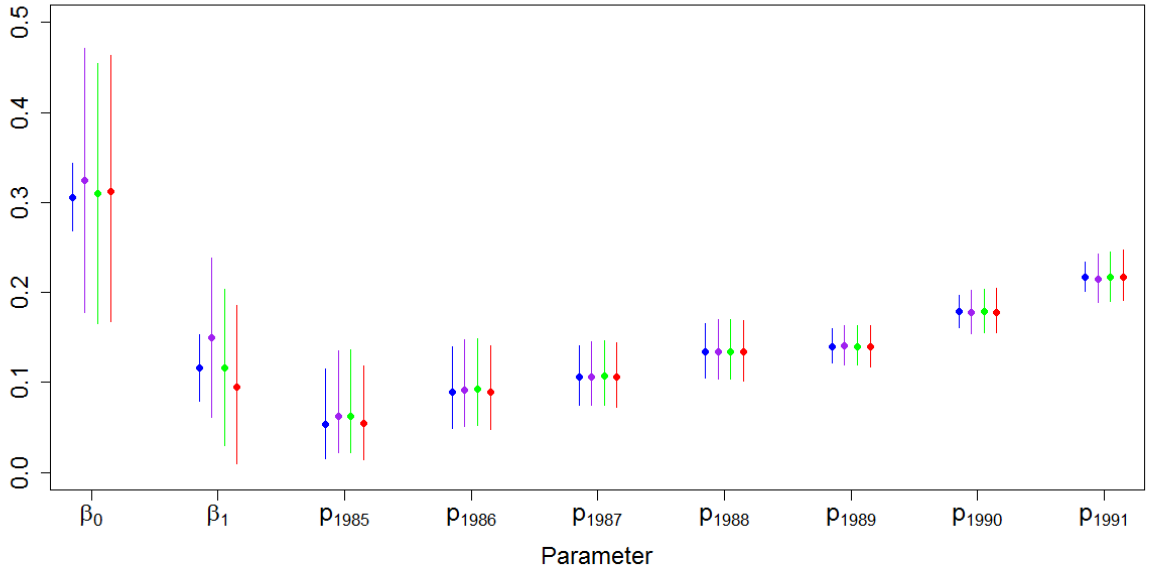


Figure 3.14: *Comparison of Parameter Estimates for a Subset of Cliff Swallows Data*

MCMC (red), MFVB (blue), correlation corrected MFVB (purple), and hybrid MFVB (green) posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters for the first 8 years of cliff swallows data. Parameter labels are located on the x-axis.

The hybrid algorithm and the original MFVB algorithm were also the most computationally efficient, taking a little more than a half hour to converge. Meanwhile, the correlation corrected MFVB algorithm and MCMC algorithm required 12 hours and 23 minutes and 7 hours and 9 minutes, respectively. As in the simulation results from Section 3.3, the hybrid algorithm shows comparable accuracy to the correlation corrected MFVB algorithm for a fraction of the run time. The correlation corrected algorithm also demonstrates its inability to handle large mark-recapture data sets, requiring even more time than the MCMC algorithm to converge. For this reason, only the hybrid and original MFVB algorithms will be used to analyze the full cliff swallows data set.

Figure 3.15 displays the results generated from the original and hybrid MFVB algorithm when applying the CJS model extended to include an effect of weight on survival to the full 29 years of cliff swallows data. The two algorithms generate very similar posterior means, although the posterior variance is higher for the survival parameters for the hybrid MFVB results. This is consistent with the simulation study and analysis of the first 8 years of cliff swallows data. Additionally, the aforementioned previous results demonstrated that this higher posterior variance is more consistent with the true posterior distribution (represented by posterior samples generated via an MCMC algorithm).

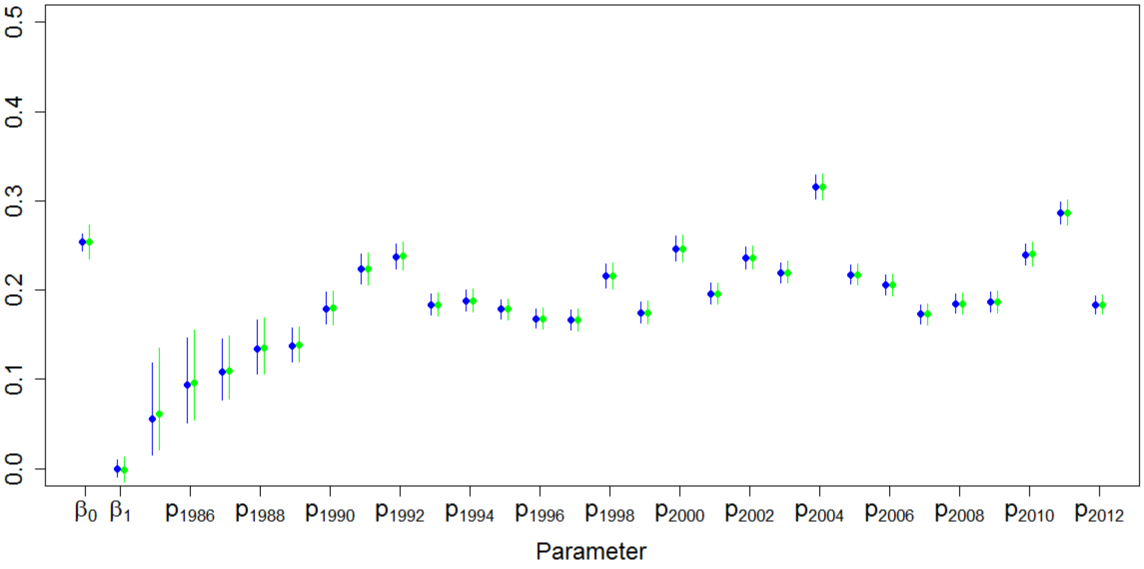


Figure 3.15: *Comparison of Parameter Estimates for Cliff Swallows Data*

MFVB (blue) and hybrid MFVB (green) posterior means (points) and 95% credible intervals (vertical lines) for capture and survival parameters for the entire 29 years of cliff swallows data. Parameter labels are located on the x-axis.

Interpreting these results, it does not appear that the weight of a cliff swallow has a significant effect on its survival as the hybrid algorithm estimated that β_1 has a posterior mean of -0.001 with a 95% credible interval of $(-0.015, 0.013)$. Additionally, an average cliff swallow has a survival probability of around 56% from one year to another. Also, the capture probabilities vary from year to year, ranging from a high of 32% in 2004 to a low of 6% in 1985.

This algorithm required over 22 days to converge. Although this result is somewhat disappointing and I had hoped for a faster, more efficient algorithm, using the traditional MCMC approach would have taken even longer to converge (when analyzing the smaller subset of cliff swallows data, the MCMC algorithm converged over 14 times more slowly than my MFVB algorithm, suggesting that the MCMC algorithm would have required almost a year to provide the same information). As I will discuss in more detail in Section 3.5, the reason my MFVB algorithm does not scale as well as I had hoped is the same reason that the MCMC algorithm has difficulties: I still need to impute every single missing covariate on each iteration of the algorithm.

3.5 Discussion

In conclusion, the MFVB algorithm introduced in Section 2.2 results in optimal variational distributions that converge much more quickly than an MCMC approach. However, the optimal variational distributions from the MFVB algorithm are an approximation of the true posterior distribution and significantly underestimate the posterior variance, as my simulation studies have shown. The primary reason for the inaccuracy with regard to the posterior variance is that the survival and capture parameters are often highly correlated. Section 3.1 demonstrated that a similar phenomenon could be observed by simulating data and fitting a mixed effects model via the MFVB framework. After deriving a solution to the mixed model accuracy problem, I applied a similar solution to the mark-recapture model. Allowing correlation between the survival and capture parameters in the target approximate posterior distribution resulted in the variational Bayesian algorithm presented in Section 3.2. This algorithm produced optimal variational densities that converged to approximate posterior distributions that are much closer to the true posterior distributions, although the multidimensional numerical optimization procedure required causes very slow convergence (even slower than the chains produced by an MCMC algorithm in some situations).

The hybrid algorithm introduced in Section 3.3 attempts to achieve the accuracy of the correlation corrected MFVB algorithm while maintaining the convergence speed of the optimal variational densities from the original MFVB algorithm. The hybrid algorithm utilizes the original MFVB algorithm until convergence of the optimal variational densities is achieved, then computes a single iteration of the correlation corrected MFVB algorithm as a correction for the underestimated posterior variance. I have found that this works very well in practice, as the original MFVB can properly estimate the posterior means of the nuisance parameters (missing covariates, $\boldsymbol{\Delta}$, and τ) just as well as the correlation corrected MFVB algorithm in a fraction of the time. If the nuisance parameter estimates from the original MFVB algorithm are similar to the nuisance parameter estimates that the correlation corrected MFVB algorithm produces, then the final iteration of the hybrid algorithm will result in an approximate posterior distribution similar to what the correlation corrected MFVB algorithm would have produced. Although the hybrid algorithm lacks a rigorous mathematical justification, I have shown via simulation study and the cliff swallows data that this admittedly *ad hoc* method performs well.

While I have demonstrated that the hybrid algorithm results in optimal variational densities that converge much faster than an MCMC approach with comparable accuracy, my MFVB algorithms still do not scale very well to extremely large data sets, such as the cliff swallows data. The optimal variational densities from the hybrid algorithm required 22 days to reach convergence when analyzing the full 29 years of cliff swallows data, and while this is an improvement over an MCMC approach, I would like to do better. Unfortunately, the MFVB algorithms will scale just as poorly as the MCMC algorithms, due to the fact that they do not truly address the reason that MCMC methods struggle with large data sets that feature individual, continuous, time varying covariates: the imputation of missing covariates.

In my MFVB approaches, I am still required to impute every single missing covariate at every iteration. If these updates only required a simple, closed form update, this would not be much of an issue. However, Algorithm 1 reveals that the missing covariate updates require numerical optimization due to the fact that their approximate posterior distribution is approximated via Laplace's method. While this numerical optimization is one dimensional and therefore not overly resource intensive, it does need to occur on each iteration of the algorithm and will scale poorly as the data sets increase in size, with respect to both n and T, as this will increase the number of missing covariates.

For an algorithm to more successfully address the large data problem, I must devise an approach that reduces the computational burden associated with imputing missing covariates. This is precisely what I attempt to do in Chapter 4, and, although my approach in that section does not involve the use of variational Bayesian methods, I have demonstrated in this section that variational Bayesian methods can be effective in reducing convergence time when applied mark-recapture models. The decision to employ variational Bayesian methods as an alternative to MCMC for mark-recapture models similar to the modified CJS model we've been examining will involve considering the trade-off between improved convergence speed and the need to mathematically derive the updates to the variational Bayesian algorithm. While the algorithms presented in Sections 2.2 and 3.2 can be relatively easily modified to apply to other CJS extensions, the process will never be as easy as modifying the model file of a general MCMC software package such as JAGS, BUGS, or STAN. Nevertheless, if convergence speed is of primary concern, the variational Bayesian techniques outlined in this section will prove useful.

Chapter 4 Truncated CJS

While implementing a variational Bayesian algorithm as an alternative to MCMC does result in improved convergence speed, it does not address the primary reason that fitting CJS models extended to incorporate individual, continuous, time-varying covariates to very large data sets scales so poorly. To truly address this issue, I must find a way to spend fewer resources imputing missing covariates. The repeated updating of these values makes the traditional MCMC approach unfeasible for extremely large data sets, especially when there are many capture occasions and individuals are short lived or capture rates are low. I address this in this chapter by introducing the truncated CJS model.

By focusing on the data that have the most influence on the parameter estimates, the truncated CJS model allows us to decrease the number of missing covariates I am required to update on each MCMC iteration. A tuning parameter, k, is also included, allowing researchers to have control over the trade-off between efficiency and increased parameter certainty.

4.1 Introduction to Truncated CJS Model

Notation

In addition to the notation introduced in Section 1.2, I must introduce some additional notation in order to define the truncated CJS model:

$r_i =$ number of times individual i was released,
not including releases on the final occasion

$t_{i,1}, \ldots, t_{i,r_i} =$ occasions on which individual i was released,
not including the final occasion

$Y_{i,j} =$ number occasions between release j of individual i and
the next recapture or -1 if individual i is not recaptured after release j

$Y_{i,j}^{(k)} =$ number occasions between release j of individual i and
the next recapture if the next recapture occurs k occasions
after release or sooner, -1 otherwise

Alternative CJS Likelihood

Recall the likelihood associated with the CJS model introduced in Section 1.3, in which individual capture histories were assigned probabilities and these probabilities were function of survival and capture parameters. The truncated CJS model I will introduce in the next section extends from an equivalent definition of the CJS likelihood function constructed by considering separate likelihood contributions for each release of each individual instead of considering the contributions for the capture histories of each individual as a whole. Let r_i denote the number of releases and $t_{i,1}, \dots, t_{i,r_i}$ the occasions of release for individual i, not including the final occasion. The likelihood in equation (1.1) can then be written as

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\phi}|\boldsymbol{\Omega}) = \prod_{i=1}^{n} \prod_{j=1}^{r_i} Pr(Y_{i,j}|t_{i,j}) \tag{4.1}$$

where

$$Y_{i,j} = \begin{cases} t_{i,j+1} - t_{i,j} & \text{if} \quad j < r_i \\ -1 & \text{if} \quad j = r_i \end{cases} . \tag{4.2}$$

Heuristically, $Y_{i,j}$ is equal to the number of occasions between release and recapture or -1 if the individual not recaptured after $t_{i,j}$. If $T = 3$, then there are only five possible pairs of release and recapture times, and the probabilities assigned to these pairs are given in Table 4.1. Note that each entry in Table 1.1 is a product of these new probabilities. More generally, there are $\sum_{t=2}^{T} t = \frac{T(T+1)}{2} - 1$ possible pairs of release and recapture times for an experiment with T occasions and the new probabilities are

$$P(Y_{i,j} = y|t_{i,j}) = \begin{cases} \phi_{i,t_{i,j}} p_{i,t_{i,j}+1} & y = 1 \\ \phi_{i,t_{i,j}} \phi_{i,t_{i,j}+1} \phi_{t_{i,j}+2} (1 - p_{t_{i,j}+1}) p_{t_{i,j}+2} & y = 2 \\ \phi_{i,t_{i,j}} \cdots \phi_{t_{i,j}+y-1} (1 - p_{t_{i,j}+1}) \cdots (1 - p_{t_{i,j}+y-1}) p_{t_{i,j}+y} & y \geq 3 \\ \chi_{i,t_{i,j}} & y = -1. \end{cases} \tag{4.3}$$

Note that $\chi_{i,t}$, the probability that individual i is not recaptured after occasion t, can also be defined as $\chi_{i,t} = 1 - \sum_{s=t+1}^{T} P(Y_{i,j} = s|t_{i,j})$ by simple complements.

Table 4.1: *Possible probabilities assigned to individual i's capture history*

Release	Time to	Probability
Occasion (t_i)	Recapture ($Y_{i,t}$)	
1	2	$\phi_{i,1}p_{i,2}$
1	3	$\phi_{i,1}(1-p_{i,2})\phi_{i,2}p_{i,3}$
1	-1	χ_{i1}
2	3	$\phi_{i,2}p_{i,3}$
2	-1	$\chi_{i,2}$

4.2 Truncated CJS Likelihood

I define the new likelihood function associated with the truncated CJS model by truncating the release specific times to recapture, $Y_{i,j}$, defined in equation (4.2). Specifically, I define truncated recapture times for some pre-determined k representing the maximum number of occasions I consider until an individual is recaptured. The truncated recapture times, denoted by $Y_{i,t}^{(k)}$, are then defined such that $Y_{i,t}^{(k)} = Y_{i,t}$ if $Y_{i,t} \leq k$ and $Y_{i,t} = -1$ otherwise. Probabilities assigned to the events $Y_{i,t}^{(k)} = 1, \ldots, Y_{i,t}^{(k)} = k$ are defined exactly as in equation (4.3). However, the event $Y_{i,t}^{(k)} = -1$, now combines the events that an individual is recaptured more than k occasions after it is released or not at all. The probability assigned to this event is most easily computed as one minus the sum of the probabilities assigned to the other events so that

$$P(Y_{i,t}^{(k)} = y|t_{i,j}) = \begin{cases} \phi_{i,t_{i,j}}p_{i,t_{i,j}+1} & y = 1 \\ \phi_{i,t_{i,j}}\phi_{i,t_{i,j}+1}\phi_{t_{i,j}+2}(1-p_{t_{i,j}+1})p_{t_{i,j}+2} & y = 2 \\ \phi_{i,t_{i,j}}\cdots\phi_{t_{i,j}+y-1}(1-p_{t_{i,j}+1})\cdots(1-p_{t_{i,j}+y-1})p_{t_{i,j}+y} & y = 3, \ldots, k \\ 1 - \sum_{s=1}^{k} P(Y_{i,t}^{(k)} = s|t_{i,j}) & y = -1 \end{cases}$$

The new likelihood function is then defined as

$$\mathcal{L}^{(k)}(\boldsymbol{p}, \boldsymbol{\phi}|\boldsymbol{\Omega}) = \prod_{i=1}^{n}\prod_{j=1}^{r_i} Pr(Y_{i,j}^{(k)}|t_{i,j})$$

after simply replacing $Pr(Y_{i,j}|t_{i,j})$ with $P(Y_{i,t}^{(k)} = y|t_{i,j})$.

Estimates of the capture and survival parameters (and other parameters if, for example, a covariate is being modeled) can then be obtained by maximizing $\mathcal{L}^{(k)}(\boldsymbol{p}, \boldsymbol{\phi}|\boldsymbol{\Omega})$ to obtain MLEs or by constructing a posterior distribution based on this likelihood in the Bayesian context. The truncation parameter, k, can be seen as a tuning parameter. The smaller k is, the faster the parameter estimation will take place, but more data will be discarded and the credible regions or confidence intervals will be larger. It is immediate that setting $k = T$ results in the likelihood from the CJS model as defined in Section 1.4.

Note that an individual capture history may be truncated more than once and that truncation may occur before an individuals last capture. For example, the capture history $\omega_i = 100001000$ would be truncated twice when fitting the truncated CJS model with $k = 3$, contributing both $P(Y_{i,1}^{(3)} = -1)$ and $P(Y_{i,6}^{(3)} = -1)$ to the likelihood.

In the presence of extremely large data sets (with respect to n and/or T), fitting the truncated CJS model can lead to significant improvements in the computational efficiency of the parameter estimation algorithms. Truncation is especially effective when capture probabilities are high and survival probabilities are low. Suppose, for example, that individual i is marked on the first occasion of a study with $T = 30$ capture occasions and never seen again. The likelihood contribution for this individual is simply $\chi_{i,1}$. This is a function of $\phi_{i,1}$, ..., $\phi_{i,29}$, $p_{i,2}$, ..., $p_{i,30}$. However, it is unlikely that the individual lived to the end of the study and remained uncaptured if either the survival probabilities are low or the capture probabilities are high. The truncated CJS model allows us to discard the least influential data points in this individual's capture history (in this case, the capture occasions late in the study after this individual is very likely deceased) while still producing unbiased estimates of capture and survival parameters. The ability to discard weakly influential data points can greatly reduce the computational burden of MCMC algorithms in the presence of continuous, individual, time-varying covariates, as I will demonstrate in Sections 4.4 and 4.5. First, I quantify how much precision is lost when truncating capture histories in Section 4.3.

4.3 Accuracy and Precision of the Truncated CJS Model

The truncated CJS model aims to improve algorithmic efficiency by focusing only on the most influential data. However, this improvement in efficiency comes at the cost of ignoring less influential data, which will decrease the precision of parameter estimates. In this section, I analytically quantify the precision that is lost when capture histories are truncated. In Section 4.4, I examine the performance of the truncated CJS model via a simulation study.

To make the analytical calculation tractable, I consider a simple model in which all n individuals in the sample are marked and released at the beginning of the study.

Additionally, the capture probability, p, and survival probability, ϕ, are assumed to be common for all individuals in the study and do not vary across capture occasions. Let m_t represent the number of individuals first recaptured on capture occasion t, m_0 represent the number of individuals never recaptured, and k represent the duration of the study. Note that because all individuals are marked and released at the beginning of the study and we are only concerned with the first recapture, k is analogous to the truncation parameter from the truncated CJS model rather than the study duration, T. The likelihood for this simplified model is:

$$\mathcal{L}(\phi, p) = \left(\prod_{t=1}^{k} P_t^{m_t} \right) \times P_0^{m_0}$$

where

$$P_t = \phi^t (1-p)^{t-1} p$$

is the probabilty that an individual is recaptured after t occasions and

$$P_0 = 1 - \sum_{t=1}^{k} P_t$$

is the probability that an individual is never recaptured. I can estimate the survival and capture parameters from this simplified CJS model using maximum likelihood techniques. Furthermore, because of the model's simplicity, I am able to compute both the Fisher information matrix and its inverse (the asymptotic variance-covariance matrix of the maximum likelihood estimator).

Having the asymptotic distribution of the maximum likelihood estimator in closed form allows us to analytically determine the effect of truncation on precision. To evaluate the degree to which k affects the precision of both the survival and capture parameter estimates simultaneously, I will use the Kullback-Liebler distance between the asymptotic multivariate normal distribution of the MLEs at a fixed value of k and the asymptotic multivariate normal distribution of the MLEs when $k \to \infty$. I chose to use the Kullback-Liebler distance as the metric because the results can be interpreted as the amount of information lost when approximating one distribution with another (Burnham and Anderson, 1998, pg. 51).

Figures 4.1 and 4.2 display the analytically calculated Kullback-Liebler distances between the asymptotic multivariate normal distribution of the MLEs at 6 different values of k and the asymptotic multivariate normal distribution of the MLEs when

$k \to \infty$. Additionally, Figure 4.1 presents Kullback-Liebler distances at 3 different values of p (0.5, 0.7, and 0.9) and a single value of ϕ (0.7), while Figure 4.2 presents Kullback-Liebler distances at 3 different values of ϕ (0.5, 0.7, and 0.9) and a single value of p (0.7). These figures show that, when fitting the simplified CJS model described above to data in which the true values of p and ϕ are between 0.5 and 0.9, a study duration, k, of 5 capture occasions or more results in asymptotic maximum likelihood estimates of ϕ and p that are nearly indistinguishable from those in which $k \to \infty$. Additionally, I found that asymptotic maximum likelihood estimates from smaller study durations are more similar to to the asymptotic maximum likelihood estimates when $k \to \infty$ when the true, underlying values of ϕ are low and p are high. Conversely, when the true values of ϕ are high and p were are low, the Kullback-Liebler distances between the asymptotic distribution of the MLEs of low, fixed values of k and $k \to \infty$ are the largest.

The results presented in Figures 4.1 and 4.2 are intuitive. When individuals are long lived and rarely captured more capture occasions will be required to accurately estimate the capture and survival probabilities. The analytical results presented in this section apply to a very simple mark-recapture model. However, the information obtained is applicable to the more complex truncation of the CJS model, as k, the length of the study in the simplified CJS model presented in this section, still represents the truncation of capture histories. The simulation study presented in Section 4.4 and the example data set analyzed in Section 4.5 came to similar conclusions, as a truncation parameter (k) of 5 was sufficient for the posterior distributions to closely resemble those obtained when fitting a model with no truncation of capture histories in both scenarios.

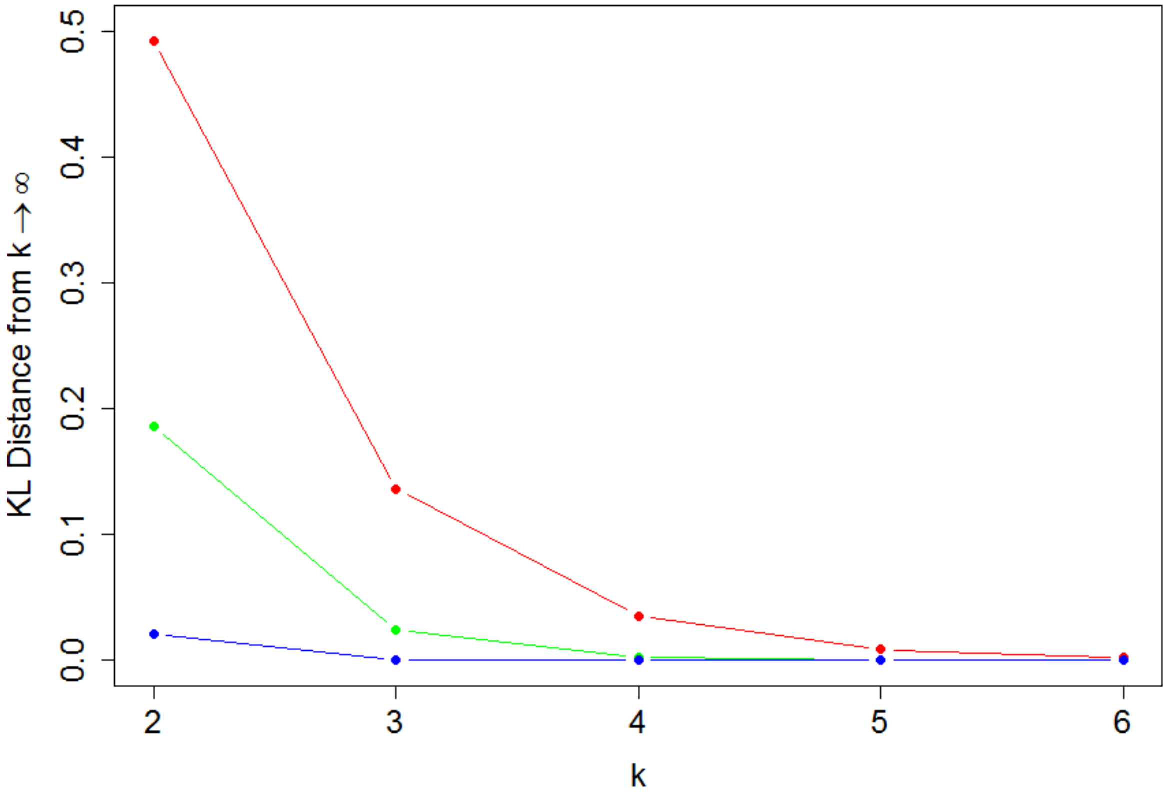


Figure 4.1: *KL Distances between MLEs at Different Values of k and p*

This figure displays the Kullback-Liebler distances between the asymptotic multivariate normal distributions of the MLEs at 6 different values of k with the asymptotic multivariate normal distributions of the MLEs as $k \to \infty$. I fix ϕ at 0.7 and present the KL distances for 3 different values of p: $p = 0.5$ (in red), $p = 0.7$ (in green), and $p = 0.9$ (in blue).

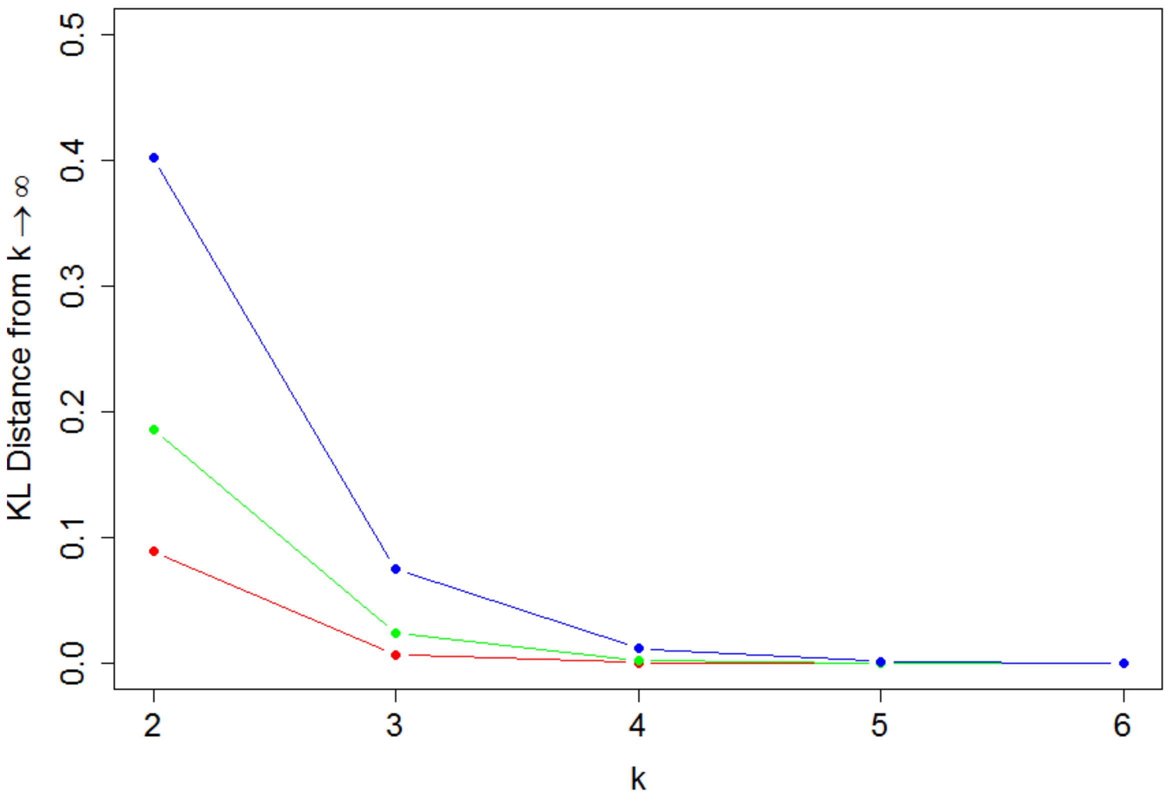


Figure 4.2: *KL Distances between MLEs at Different Values of k and ϕ*

This figure displays the Kullback-Liebler distances between the asymptotic multivariate normal distributions of the MLEs at 6 different values of k with the asymptotic multivariate normal distributions of the MLEs as $k \to \infty$. I fix p at 0.7 and present the KL distances for 3 different values of ϕ: $\phi = 0.5$ (in red), $\phi = 0.7$ (in green), and $\phi = 0.9$ (in blue).

4.4 Simulation Study

To ensure that the parameter estimates generated from the truncated CJS model are unbiased, confirm the improved efficiency of the truncated CJS model, and determine how changing k affects precision, I analyzed data sets simulated with known parameter values and assessed the performance of the truncated CJS model in the presence of continuous, time varying covariates. Note that the truncated CJS model may be extended to incorporate covariates in the same manner as the traditional CJS model was in Section 1.4, as both likelihoods are functions of the same capture and survival parameters, which can be modeled as a function of covariates via a logit link function.

The truncated CJS model also does not impose any specific model for the missing covariates, and therefore the process described by Bonner and Schwarz (2006) that was introduced in Section 1.4 may also be implemented here.

I generated 100 data sets from two simulation scenarios with $n = 600$ individuals in each in order to assess the degree to which fitting the truncated CJS model will result in increased algorithmic efficiency in the presence of continuous, time-varying covariates. The first 100 data sets consisted of $T = 15$ capture occasions while the second group of 100 data sets consisted of $T = 20$ capture occasions, to demonstrate how the efficiency gains from the truncated CJS model improve as the number of capture occasions increases. Individual, time-varying covariates were included, with the initial covariate value generated from the uniform distribution on $(-0.5, 0.5)$. After the first capture, the covariate values followed the model in Equation 1.2 with $\Delta_t = 0.4$ for all t and $\tau = 1$. I included covariate effects on both the capture and survival probabilities, with $\beta_{0,p} = -0.4$, $\beta_{1,p} = 1$, $\beta_{0,\phi} = 1$, and $\beta_{1,\phi} = -0.3$. These parameters lead to an individual with a covariate value of 1 having a survival probability of around 67% and a capture probability of 65%. An individual with a covariate value of 0 would have a survival probability of 73% with a capture probability of 40% and an individual with a covariate value of -1 would have a survival probability of 79% and a capture probability of 20%.

For each scenario, the truncated CJS model was implemented with $k = 2$, $k = 3$, $k = 5$, and $k = T$ (which replicates the original CJS model) using JAGS version 4.2.0, a statistical software package that performs Gibbs sampling (Plummer, 2003). When $T = 15$, I generated 3 Markov chains of length $10,000$ for each data set, discarding the first half as burn-in. When $T = 20$, more iterations were required to reach convergence and 3 chains each of length $20,000$ were generated.

Table 4.2 shows the estimated percent bias and average relative standard error of the four capture and survival parameters ($\beta_{0,p}$, $\beta_{1,p}$, $\beta_{0,\phi}$, and $\beta_{1,\phi}$), in addition to the average run time required to fit the model using each algorithm at the four different values of k when $T = 15$. I define the percent bias and relative standard error as the bias and standard error divided by the absolute value of the true parameter value used to generate the data multiplied by 100 (e.g. the percent bias for $\beta_{0,p}$ is given by $(\mathrm{E}[\widehat{\beta_{0,p}}] - \beta_{0,p})/|\beta_{0,p}| \times 100$). Observe that the estimated percent bias never exceeds $\pm 7\%$ for any parameter at any value of k, but that the estimated percent bias and average relative standard error of each parameter do get larger as k decreases.

Note that while the relative standard error and percent bias estimates were highest when $k = 2$, the algorithm ran in less than one third of the time required to fit the non-truncated CJS model. Likewise, when $k = 3$ the MCMC algorithm required less than half the time required to fit the non-truncated CJS model. Additionally, models with lower values of k result in MCMC algorithms that mix better and converge to the posterior distribution more quickly, due to the fact that fewer missing covariates occurring well after an individual's last capture need to be imputed. This property is exhibited by the effective samples per second column in Table 4.2, which I define as the sum of the effective sample sizes for the posterior samples of $\beta_{0,p}$, $\beta_{1,p}$, $\beta_{0,\phi}$, and $\beta_{1,\phi}$ divided by the run time (in seconds). Note that effective sample size is a measure of posterior sample size that has been corrected to account for the autocorrelation present in MCMC chains. This measure of algorithm efficiency shows that algorithms fitting models with smaller values of k not only run faster when the chains are of equal length, but will also generate more effective samples given a fixed run time.

Table 4.3 contains similar results for simulations in which $T = 20$. Note that the efficiency improvements that the truncated CJS model demonstrated over the non-truncated CJS model become even more pronounced for data sets with more capture occasions.

Figures 4.3 and 4.4 compare the posterior densities from all four values of k for each β_j for a single simulated data set with $T = 15$ and $T = 20$. As seen in the cumulative results, the parameter estimates are more uncertain as k decreases. All the densities, however, are roughly centered around the same estimate.

Table 4.2: *Run times, estimated % bias, average relative standard errors, and effective samples per second when fitting the truncated CJS model (at 4 different values of k) to a data set with $n = 600$ and $T = 15$ capture occasions.*

	$\beta_{0,p}$		$\beta_{1,p}$		$\beta_{0,\phi}$		$\beta_{1,\phi}$			
k	Estimated % Bias	Average Relative SE	Estimated % Bias	Average Relative SE	Estimated % Bias	Average Relative SE	Estimated % Bias	Average Relative SE	Run Time (minutes)	Effective Samples per Second
2	-3.4%	34.8%	6.3%	13.0%	3.7%	16.2%	-6.6%	24.2%	17.8	4.4
3	-4.3%	29.3%	5.2%	12.1%	2.4%	11.3%	-4.6%	20.3%	24.4	2.5
5	-4.4%	26.0%	5.2%	11.4%	1.7%	8.7%	-3.7%	18.0%	38.1	1.9
T	-2.0%	24.8%	4.4%	10.9%	0.4%	7.7%	-1.5%	16.9%	53.6	1.2

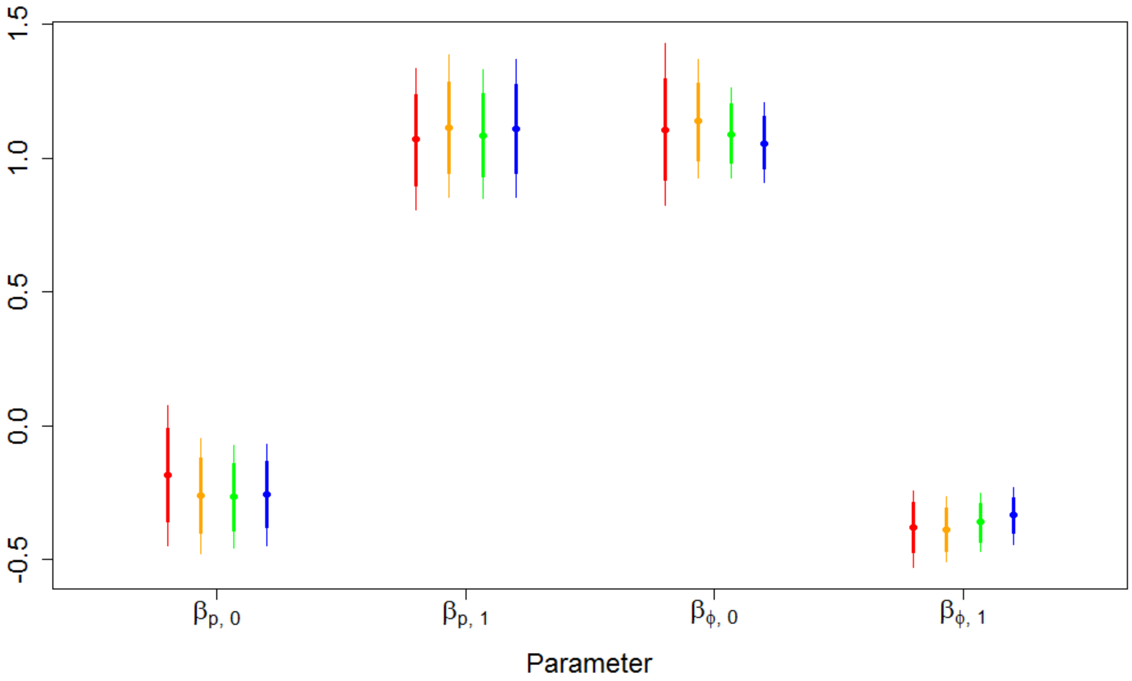


Figure 4.3: *Parameter Estimates from the Truncated CJS Model at Different Values of k for Simulated Data with $T = 15$*

Posterior means (points), 80% credible intervals (thick vertical lines), and 95% credible intervals (thin vertical lines) for capture and survival parameters estimated from a single simulated data set with $n = 600$ and $T = 15$ capture occasions. The results from the truncated CJS model with $k = 2$ (red), $k = 3$ (orange), $k = 5$ (green), and the full CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

Table 4.3: *Run times, estimated % bias, average relative standard errors, and effective samples per second when fitting the truncated CJS model (at 4 different values of k) to a data set with $n = 600$ and $T = 20$ capture occasions.*

	$\beta_{0,p}$		$\beta_{1,p}$		$\beta_{0,\phi}$		$\beta_{1,\phi}$			
k	Estimated % Bias	Average Relative SE	Estimated % Bias	Average Relative SE	Estimated % Bias	Average Relative SE	Estimated % Bias	Average Relative SE	Run Time (minutes)	Effective Samples per Second
2	-2.1%	40.4%	8.1%	15.0%	5.6%	19.3%	-7.8%	28.9%	65.8	1.5
3	-4.6%	33.8%	7.5%	13.7%	4.6%	13.5%	-6.4%	24.1%	110.7	0.7
5	-7.7%	29.9%	10.1%	12.9%	3.2%	10.3%	-5.0%	21.2%	184.1	0.4
T	-11.8%	28.6%	13.2%	12.0%	1.2%	9.1%	-3.8%	19.8%	320.0	0.2

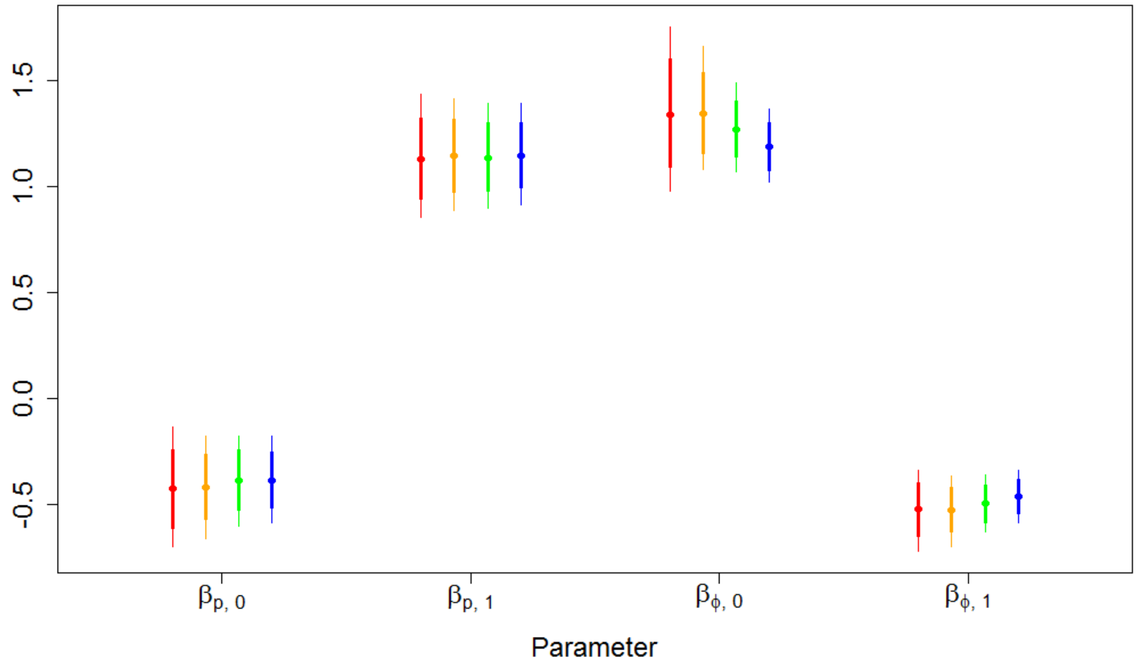


Figure 4.4: *Parameter Estimates from the Truncated CJS Model at Different Values of k for Simulated Data with $T = 20$*

Posterior means (points), 80% credible intervals (thick vertical lines), and 95% credible intervals (thin vertical lines) for capture and survival parameters from a single simulated data set with $n = 600$ and $T = 20$ capture occasions. The results from the truncated CJS model with $k = 2$ (red), $k = 3$ (orange), $k = 5$ (green), and the full CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

4.5 Application to Cliff Swallows

Recall the data set consisting of records from 164,621 cliff swallows observed over a 29 year period introduced in Section 1.5. Fitting the truncated CJS model is particularly useful for improving the computational efficiency of the fitting algorithm for this data set. This is because individuals who are only observed early in the study will, using the non-truncated CJS model with continuous covariates, continue to have their covariates imputed on many capture occasions after their last sighting when many individuals are likely deceased. For example, fitting the non-truncated CJS model with continuous covariates with this data set requires imputing $1,968,151$ covariate values each iteration of an MCMC algorithm. However, fitting the truncated

CJS algorithm with $k = 5$ reduces this number to $742,518$, while $k = 3$ reduces this number to $478,806$. This is the primary reason that computational efficiency is improved when fitting the truncated CJS algorithm.

The model I have fit includes time-varying intercepts for each year and an effect of weight on both capture and survival. Note that the model considered in Chapters 2 and 3 did not consider an effect of weight on capture or survival probabilities with time-varying intercepts (this led to variational Bayesian algorithms with derivations that were easier to follow). Weights were standardized before analysis. The first 8 years of data was analyzed initially to assess how well models using different values of k performed compared to the non-truncated CJS model. The truncated CJS model was fit to this subset of data with $k = 2$, $k = 3$, and $k = 5$. I generated 3 Markov chains each of length $50,000$ from the Gibbs sampler, discarding the first $2,000$ as burn-in. JAGS version 4.2.0 was used to generate the posterior samples (Plummer, 2003), which were then processed through R (R Core Team, 2014) to create summaries, tables, and graphics.

Figures 4.5 and 4.6 compare the estimated posterior distributions of parameters for the first 8 years of cliff swallows data. The posterior distribution plots show that when $k = 3$ or $k = 5$, the estimated posterior distributions match quite closely to the estimated posterior distribution from the non-truncated CJS model with continuous covariates. When $k = 2$, it is clear that the estimated marginal posterior distributions of the capture parameters are drastically overestimated while the survival parameters are underestimated (when compared to the estimated posterior distribution obtained when fitting the non-truncated CJS model). I did not see this systematic bias in the simulation results presented in Section 4.4, indicating that this data deviates from the assumptions made in the modeling process in some way. One possible deviation would be the presence of temporary emigration (i.e. individuals leave the study population but eventually return). I explore the effects of temporary emigration further in Appendix A.4 and find similar systematic bias with respect to the choice of k. Table 4.4 displays the performance gains associated with lower values of k, detailing the run time in hours and the effective samples per second (the sum of the effective sample sizes for all β parameters divided by the run time in seconds).

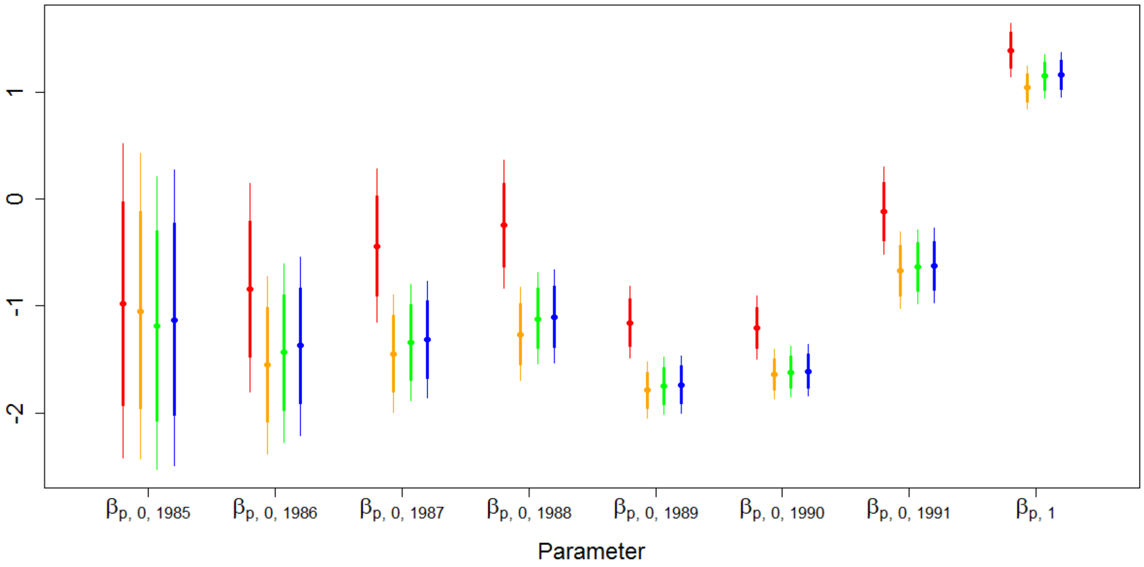


Figure 4.5: *Estimates of Capture Parameters from the Truncated CJS Model for 8 Years of Cliff Swallows Data*

Posterior means (points), 80% credible intervals (thick vertical lines), and 95% credible intervals (thin vertical lines) for the capture parameters estimated from the first 8 years of cliff swallows data. The results from the truncated CJS model with $k = 2$ (red), $k = 3$ (orange), $k = 5$ (green), and the full CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

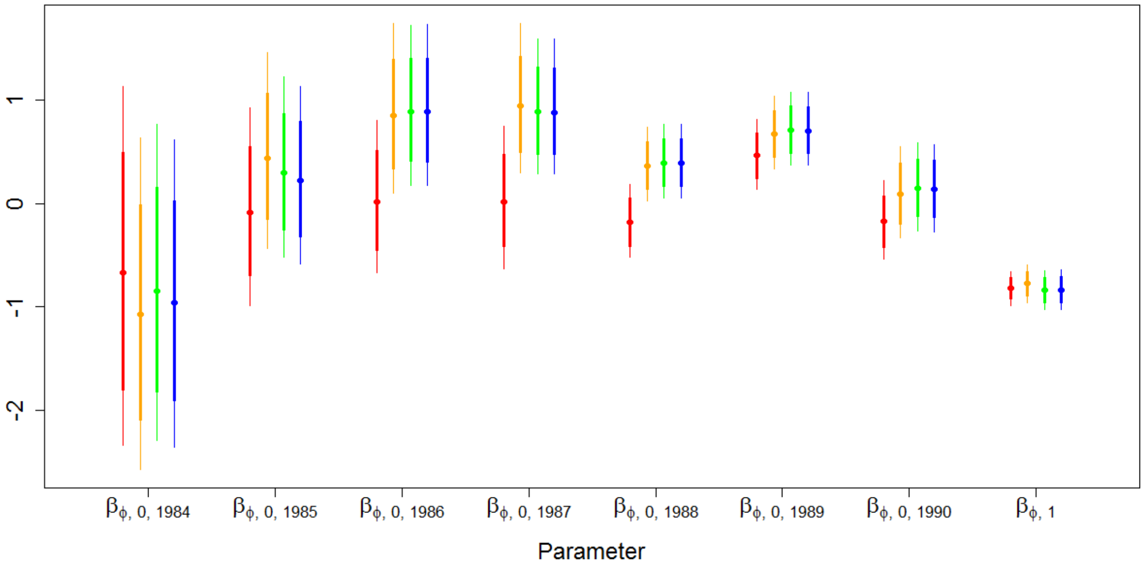


Figure 4.6: *Estimates of Survival Parameters from the Truncated CJS Model for 8 Years of Cliff Swallows Data*

Posterior means (points), 80% credible intervals (thick vertical lines), and 95% credible intervals (thin vertical lines) for the survival parameters estimated from the first 8 years of cliff swallows data. The results from the truncated CJS model with $k = 2$ (red), $k = 3$ (orange), $k = 5$ (green), and the full CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

Table 4.4: *Run times and effective samples per second when fitting the truncated CJS model (at 4 different values of k) to the first 8 years of the cliff swallow data.*

k	Run Time (hours)	Effective Samples per Second
2	12.5	7.6
3	15.9	4.9
5	18.5	0.8
T	18.6	1.8

Finally, I analyzed the entire 29 year data set containing records from the 164,621 birds with at least one recorded weight. Due to the size of the data set, it is no longer computationally feasible to fit the non-truncated CJS model. Additionally, due to the poor accuracy of the estimates obtained when fitting the truncated CJS model with $k = 2$, I only fit the truncated CJS models with $k = 3$ and $k = 5$ to the full cliff swallows data set. The estimated capture parameters are presented in Figure 4.7 while the estimated survival parameters are presented in Figure 4.8. Lastly, even when fitting the truncated CJS model with $k = 3$, the MCMC algorithm required roughly 156 hours (nearly a week) to generate 3 Markov chains of length 50,000. When $k = 5$, the MCMC algorithm required roughly 233 hours (nearly 10 days) to generate the posterior samples. While these estimates of the posterior distribution still require many hours of computational resources to generate, this still represents a substantial improvement in computationally efficiency while producing results that are extremely similar to those that would have been generated by fitting the non-truncated CJS model to this data set.

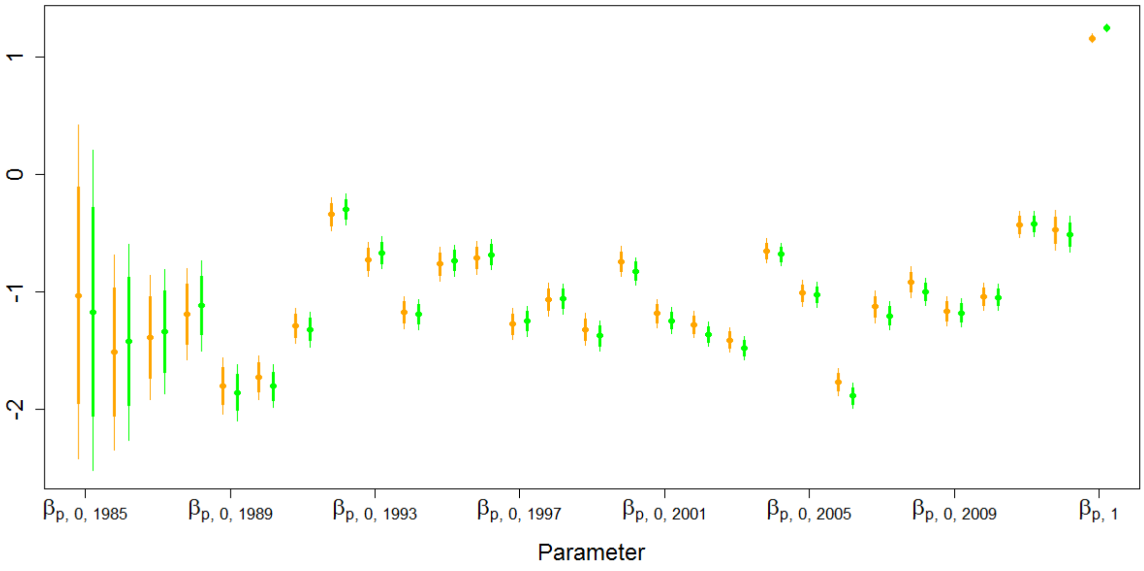


Figure 4.7: *Estimates of Capture Parameters from the Truncated CJS Model for Cliff Swallows Data*

Posterior means (points), 80% credible intervals (thick vertical lines), and 95% credible intervals (thin vertical lines) for the capture parameters estimated from the full cliff swallows data set. The results from the truncated CJS model with $k = 3$ are displayed in orange, while the results from the truncated CJS model with $k = 5$ are in green. Parameter labels are located on the x-axis.

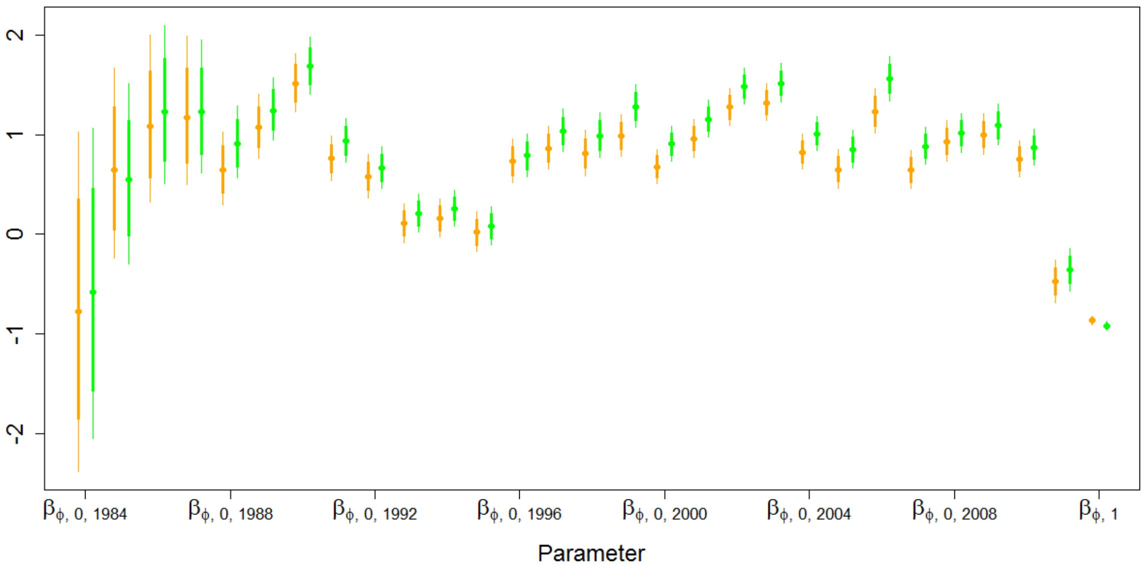


Figure 4.8: *Estimates of Survival Parameters from the Truncated CJS Model for Cliff Swallows Data*

Posterior means (points), 80% credible intervals (thick vertical lines), and 95% credible intervals (thin vertical lines) for the survival parameters estimated from the full cliff swallows data set. The results from the truncated CJS model with $k = 3$ are displayed in orange, while the results from the truncated CJS model with $k = 5$ are in green. Parameter labels are located on the x-axis.

Examining the results from fitting the truncated CJS model with $k = 5$ in more detail, the estimate of $\beta_{p,1}$ was 1.25 with a 95% credible interval of $(1.22, 1.28)$, indicating that heavier individuals are more likely to be captured. Conversely, the estimate of $\beta_{\phi,1}$ was -0.919 $(-0.964, -0.875)$, which indicates that lighter birds are more likely to survive from one capture occasion to the next. Additionally, Figure 4.7 clearly displays that the baseline capture probabilities varied significantly over time. The high point occurred in 1992, when a living individual of average weight had a 42.6% (39.4%, 45.9%) chance to be captured. A similar trend is visible in Figure 4.8, as the survival probabilities also exhibit differences over time. The high point for survival probabilities occurred in 1990, when individuals of average weight had an 84.3% (80.3%, 87.9%) chance of surviving and remaining in the population until at least 1991. Note that the results from fitting the truncated CJS model with $k = 3$ are quite similar. However, I chose to report the results with $k = 5$ because those are likely closer to the true posterior distribution of the non-truncated CJS model, based on the previously described simulation study and analysis of the first 8 years of cliff swallows data.

4.6 Discussion

In this chapter I introduced the truncated CJS model, a modification to the Cormack-Jolly-Seber model in which capture histories are truncated according to a tuning parameter. This truncation allows data sets to be fit more efficiently, especially those that contain many captured individuals, are collected over many capture occasions, or both. The increased computational efficiency associated with fitting this new model containing truncated capture histories is especially significant for models that include individual, continuous, time-varying covariates, as the need to impute covariate values when individuals are not captured exacerbates the computational challenges of fitting models to large mark-recapture data sets. The truncated CJS model does have some drawbacks, however. The increased efficiency of the model fitting algorithms, be they maximum likelihood estimation or MCMC, comes at the expense of less precise parameter estimates, as one would expect from a method that is using less data than the full, non-truncated CJS model. However, this method discards the least influential data points while still providing unbiased estimates, making the truncated CJS model an attractive option when analyzing data sets where the number of individuals and/or number of capture occasions make the use of traditional methods computationally unfeasible.

In Section 4.3 I quantified how much precision was lost by discarding the truncated data for different values of k, the truncation parameter, and for various capture and survival probabilities. In general, I found that the ratios of the standard errors from the truncated CJS model and original CJS model increased as either the survival probability increased and/or the capture probability decreased. For survival probabilities of less than 0.9 and capture probabilities greater than 0.5, I found that a truncation parameter of $k = 5$ resulted in parameter estimates with asymptotic distributions almost identical to those of the non-truncated model.

The relationship between the degree of truncation and the uncertainty associated with parameter estimates can also be seen quite clearly in my study of the truncated CJS model extended to incorporate individual, time varying covariates, via simulation in Section 4.4 and in the analysis of the cliff swallows data set in Section 4.5. Additionally, when analyzing the cliff swallows data, I noticed that there appeared to be a systematic trend with respect to k: more extreme truncation led to lower estimates of survival parameters and higher estimates of capture parameters. As I did not encounter this problem in any of my simulation scenarios, I found it likely that the cliff swallows data violated some of the assumptions made when applying the CJS model. One such possibility was the presence of temporary emigration (that is, the ability of individuals to leave the population and later return). I explored this situation via a simulation study and the results presented in Appendix A.4 are consistent with the systematic patterns, with respect to k, observed when analyzing the cliff swallows data. Other explanations, such as changing study dynamics or unmodeled behavioral effects may also be possible.

Sections 4.4 and 4.5 also demonstrated the degree to which fitting the truncated CJS model positively impacts the efficiency of the MCMC algorithm. In addition to improving the run time for a fixed posterior sample size, fitting the truncated CJS models also yielded MCMC algorithms with increased effective sample sizes per iteration and per unit time.

I believe that the truncated CJS model is a useful tool for fitting CJS-like models to mark-recapture studies that involve many individuals, take place over long periods of time, or both, particularly in the presence of individual, time-varying, continuous covariates. My method is similar to the trinomial model developed by Catchpole et al. (2008), with a few key differences. First, their method was developed for mark-recapture-recovery data in which the recovery of dead individuals assists in the estimation of survival probabilities. Second, they truncate the capture histories at $k = 1$, requiring no estimation of missing covariates whatsoever. This allows the trinomial model to avoid making modeling assumptions on the distribution of the missing covariate values. However, this also introduces difficulties when modeling the effect of individual, continuous covariates on capture probabilities, as the covariate value from the previous capture occasion must be carried forward. Bonner et al. (2010) provides a thorough comparison between the full Bayesian imputation of covariates introduced by Bonner and Schwarz (2006) and the trinomial model when fitting models that involve continuous, time-varying, covariates to mark-recapture-recovery data. By allowing the degree to which capture histories are truncated to vary via a tuning parameter, my method provides a compromise between these two methods in the presence of mark-recapture data without recoveries of dead individuals.

Lastly, I caution that an overly aggressive choice of k can inflate the uncertainty associated with parameter estimation and, in the case of data that does not strictly adhere to the assumptions of the CJS model, can result in parameter estimates quite different from those obtained from fitting the non-truncated model. In practice, I would recommend fitting the truncated CJS model at multiple values of k to smaller subsets of any large mark-recapture data set being analyzed. If strong, systematic biases are present in the parameter estimates with respect to different values of k, this likely indicates that the data set does not follow all assumptions made regarding the CJS model, such as the permanence of emigration, or the model associated with the missing covariates. The uncertainty exhibited in the parameter estimates at different values of k can also guide the choice of k, although one should be aware that this uncertainty will likely reduce considerably once the entire, large data set is included in the analysis, depending on the specifics of how the survival and capture probabilities are modeled. In addition to smaller data subsets, the analysis of data sets simulated from rough or expected parameter estimates may also be useful in selecting a value of k, though it's important to note that this will not detect data that deviates from the assumptions that underly the CJS model. More analytical work, similar to what I presented in Section 4.3 but for more complex models that include covariates, could be an area of future research.

Chapter 5 Conclusion

I have proposed two methods that reduce the computational burden associated with estimating missing covariate values when fitting the CJS model extended to include individual, continuous, time-varying covariates. In Chapter 2, I proposed a variational Bayesian algorithm in lieu of the Markov chain Monte Carlo algorithm described in Bonner and Schwarz (2006). Variational Bayesian techniques are typically much faster than MCMC approaches, although this speed comes at the cost of deriving inference from an approximate posterior distribution. This approximate posterior distribution is restricted by difficult-to-check distributional assumptions that are necessary to make optimization tractable. The particular variational Bayesian technique I have employed, mean field variational Bayes, is known to underestimate the true posterior variance (Ormerod and Wand, 2010).

Simulation studies indicated that the original mean field variational Bayesian algorithm I developed, which assumed independence between the variational distributions of the capture and survival parameters, was significantly faster than the traditional MCMC approach. However, as is expected when the mean field product restriction is not realistic, this algorithm radically underestimated the posterior variance. I attempted to correct this issue in Chapter 3, modifying the distributional assumptions underlying my variational Bayesian algorithm and allowing correlation between the survival and capture parameters in the approximate posterior distribution. This corrected algorithm resulted in approximate posterior distributions that were significantly closer to the true posterior distribution. Unfortunately, this corrected algorithm was much slower than the original VB algorithm, and converged even more slowly than the MCMC algorithm when applied to a subset of the cliff swallows data in Section 3.4. To address this issue, I created a hybrid algorithm that only uses the corrected algorithm's computationally cumbersome steps for one iteration after first converging to an approximate posterior under the originally defined VB algorithm. I found, in both simulation studies and analysis of the cliff swallows data, that this hybrid algorithm converged almost as quickly as the original VB algorithm and approximated the posterior variance just as closely as the corrected VB algorithm did.

The hybrid variational Bayesian algorithm converged much more quickly than the MCMC approach with comparable accuracy. However, all three of the variational Bayesian algorithms I implemented still do not scale well to extremely large data sets due to the same reason that MCMC algorithms do not: every missing covariate value must still be updated on every iteration of the algorithm. To demonstrate the severity of this problem, the hybrid MFVB algorithm required 22 days to converge to an approximate posterior distribution when applied to the full 29 years of cliff swallows data. Although this is better performance than I would expect from an MCMC approach, I failed to address the root cause of the computational difficulties associated with fitting this model by reducing the number of missing covariates I must estimate.

The truncated CJS model introduced in Chapter 4 aimed to reduce the number of missing covariate values that must be estimated. This is accomplished by modeling capture histories that have been truncated after an individual has not been observed for a set number of capture occasions, rather than modeling every individual's entire capture history over the whole duration of the study. Truncating the capture histories does discard some information, which results in less precise parameter estimates. However, the way in which the capture histories are truncated ensures that the least influential data is being discarded. Additionally, I have defined a tuning parameter, k, which allows investigators to determine precisely the degree to which the capture histories will be truncated. I provided guidance regarding the selection of k in this manuscript via a simplified analytical solution in Section 4.3 and a simulation study in Section 4.4. The simulation study, along with the analysis of cliff swallows data presented in Section 4.5, also demonstrated that the truncated CJS model is effective in reducing the computational burden of fitting the CJS model with continuous, time-varying, individual covariates while still producing accurate parameter estimates with only marginally less precision (depending on the chosen value of k).

I believe that the truncated CJS model has significant advantages over the existing alternative approaches to improving the efficiency of fitting the truncated CJS model with continuous, time-varying, individual covariates to very large data sets. The method most similar to the truncated CJS approach is the trinomial model developed by Catchpole et al. (2008). Like the truncated CJS model, the trinomial model reduces the number of missing covariates that must be imputed. However, this is accomplished by eliminating the need to impute missing covariates altogether by re-writing the likelihood as a product of transition probabilities and only considering transitions where covariate information is available (analogous to the truncated CJS model with $k = 1$). This model is intended to be fit to mark-recapture-recovery data, where the recovery of deceased individuals can aide in the estimation of survival parameters. Based on the performance of the truncated CJS model at very small values of k, however, the increased uncertainty in parameter estimates would quickly become untenable if the trinomial model were to be applied to mark-recapture data. Additionally, modeling capture probabilities as functions of continuous, individual covariates presents difficulties, as there are now transition probabilities where missing covariate values must be estimated in some way (typically the previous occasion's covariate value being carried forward). Lastly, Bonner et al. (2010) found that fitting the trinomial model can produce biased results with low capture probabilities and/or low sample sizes. By allowing the degree of truncation to be a tunable parameter and the estimation of some missing covariate values to be permitted, the truncated CJS model addresses these shortcomings.

Another approach, first advocated in Nichols et al. (1992) and improved upon by Langrock et al. (2013), is to discretize the continuous covariates and apply the multistate model, which can account for individual, time-varying, categorical covariates. Transforming continuous data into categorical bins does discard potentially valuable information. However, Langrock et al. (2013) pointed out that this approximation can be made arbitrarily more accurate by increasing the number of bins at the cost of computational efficiency. Unfortunately, when more than one continuous, time-varying, individual covariates are included in the model, this approach may actually become more computationally intensive than fitting the model via the MCMC approach outlined in Bonner and Schwarz (2006). The truncated CJS model does not have this problem and will outperform fitting the full, non-truncated CJS model with individual, time-varying, continuous covariates via MCMC, with regard to computational efficiency, regardless of how many covariates are included in the modeling process.

Worthington et al. (2015) applied a multiple imputation approach, first modeling the individual, continuous, time-varying covariates separately from the capture histories. After the covariate model was fit, missing covariate values were sampled from this fitted covariate model, generating multiple data sets with complete covariate information. Approximate maximum likelihood estimates were then obtained by fitting the CJS model with covariates to each of the complete data sets. Finally, these maximum likelihood estimates estimates were aggregated to account for the uncertainty in the estimation of the missing covariates. This method does not suffer from the one covariate limitation that was present in Langrock et al. (2013). However, this multiple imputation procedure must assume that the covariates are missing at random, as no capture history information is taken into account when modeling the missing covariates. The authors admitted that this is an unrealistic assumption, and while the simulation study results they presented in the paper are quite accurate, these simulations are limited. In particular, the authors only included a covariate effect on the survival probabilities. If a covariate effect is included when modeling the capture probabilities, the missing at random assumption becomes even less reasonable, as the covariates, missing when the individual is either dead or not captured, now affect both the ability to capture and the survivability of individuals. Multiple imputation techniques are known to produce biased estimates when invalid missing at random assumptions are made for simpler models (Schafer and Graham, 2002; White and Carlin, 2010; Little and Rubin, 2014), so this method could produce biased estimates when capture probabilities are influenced by a covariate. The truncated CJS method, while not as computationally efficient as the multiple imputation approach, does not make the missing at random assumption and is therefore not subject to this potential source of bias.

While I believe that the truncated CJS method possesses significant advantages over competing approaches, there is future work that can be done to improve the usefulness of this approach. The selection of k is a critical component when fitting the truncated CJS model, and more work needs to be done to prospectively assess the effect of this choice on the uncertainty of parameter estimates. Section 4.3 gave some guidance for an extremely simplistic model that contains no covariates, continuous or otherwise, and in Section 4.6 I recommended fitting the truncated CJS model with different values of k to small subsets of a larger data set, or simulated data from expected parameter values, to assess the degree of truncation that would be appropriate. However, a more sophisticated analytical solution would be ideal.

Finally, combining the two approaches discussed in this document by fitting the truncated CJS model using a variational Bayesian algorithm could lead to an even more substantial improvement in computational efficiency. Although all of the parameter estimates I presented in Chapter 4 are generated via an MCMC algorithm, this does not necessarily have to be the case. The truncated CJS model could potentially be fit through the implementation of variational Bayesian algorithm, and I demonstrated in Chapters 2 and 3 that there are significant gains in computational efficiency that could be achieved. Nevertheless, fitting the truncated CJS model via MCMC resulted in a significant gain in computational efficiency and has distinct and important advantages over competing methods.

Chapter A Appendices

A.1 Proof of Mean Field Maximization Result

The following is a proof of the result, presented in Section 2.1 that the optimal density for the K-L disance is $q^*(\theta_i) \propto \exp(E_{-\theta_i} \log p(\boldsymbol{y}, \boldsymbol{\theta}))$ under the mean field approximation, where $q(\boldsymbol{\theta})$ factorizes into $\prod_{i=1}^{M} q_i(\theta_i)$ for a partition $\{\theta_1, \theta_2, ..., \theta_M\}$ of $\boldsymbol{\theta}$ and $E_{-\theta_i}$ indicates the expectation with respect to all random variables in $\boldsymbol{\theta}$ other than θ_i. The proof comes from Ormerod and Wand (2010) with some added detail between steps.

Recall that minimizing the K-L distance is equivalent to maximizing $\mathcal{F}[q] = E_{q(\theta)}\left[\log\left(\frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right]$. The proof will start by maximizing this quantity with respect to $q_1(\theta_1)$, and all of the other optimal variational distributions of the other parameters, $q_2, ..., q_M$, will follow identical logic.

$$
\begin{aligned}
\mathcal{F}[q] =& E_{q(\theta)}\left[\log\left(\frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right]\\
=& \int q(\boldsymbol{\theta})[\log p(\boldsymbol{y},\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})]d\boldsymbol{\theta}\\
=& \int\cdots\int\prod_{i=1}^{M} q_i(\theta_i)\left[\log p(\boldsymbol{y},\boldsymbol{\theta}) - \sum_{i=1}^{M}\log q_i(\theta_i)\right]d\theta_1...d\theta_M\\
=& \int q_1(\theta_1)\left[\int\cdots\int q_2(\theta_2)...q_M(\theta_M)\log p(\boldsymbol{y},\boldsymbol{\theta})d\theta_2...d\theta_M\right]d\theta_1\\
& - \int\cdots\int\prod_{i=1}^{M} q_i(\theta_i)\log q_1(\theta_1)d\theta_1...d\theta_M - \int\cdots\int\prod_{i=1}^{M} q_i(\theta_i)\log q_2(\theta_2)d\theta_1...d\theta_M\\
& \cdots - \int\cdots\int\prod_{i=1}^{M} q_i(\theta_i)\log q_M(\theta_M)d\theta_1...d\theta_M\\
=& \int q_1(\theta_1)\left[\int\cdots\int q_2(\theta_2)...q_M(\theta_M)\log p(\boldsymbol{y},\boldsymbol{\theta})d\theta_2...d\theta_M\right]d\theta_1\\
& - \int q_1(\theta_1)\log q_1(\theta_1)d\theta_1 - \int q_2(\theta_2)\log q_2(\theta_2)d\theta_2\\
& \cdots - \int q_M(\theta_M)\log q_M(\theta_M)d\theta_M\\
=& \int q_1(\theta_1)\log\left(\frac{\exp\left(\int\cdots\int q_2(\theta_2)...q_M(\theta_M)\log p(\boldsymbol{y},\boldsymbol{\theta})d\theta_2...d\theta_M\right)}{q_1(\theta_1)}\right)d\theta_1\\
& + \text{terms not involving } q_1
\end{aligned}
$$

Next, create a valid density from $\exp\left(\int\cdots\int q_2(\theta_2)...q_M(\theta_M)\log p(\boldsymbol{y},\boldsymbol{\theta})d\theta_2...d\theta_M\right)$ by diving from it some constant C that makes the quantity integrate to 1. Call this density $\tilde{p}(\boldsymbol{y},\theta_1)$.

$$
\begin{aligned}
\mathcal{F}[q] =& \int q_1(\theta_1)\log\left(\frac{\exp\left(\int\cdots\int q_2(\theta_2)...q_M(\theta_M)\log p(\boldsymbol{y},\boldsymbol{\theta})d\theta_2...d\theta_M\right)}{Cq_1(\theta_1)}\right)d\theta_1\\
& + \log(C) + \text{terms not involving } q_1\\
=& \int q_1(\theta_1)\log\left(\frac{\tilde{p}(\boldsymbol{y},\theta_1)}{q_1(\theta_1)}\right)d\theta_1 + \text{terms not involving } q_1\\
=& E_{q_1}\left[\frac{\tilde{p}(\boldsymbol{y},\theta_1)}{q_1(\theta_1)}\right] + \text{terms not involving } q_1
\end{aligned}
$$

Next, use the fact that the K-L distance is nonnegative to derive an upper bound on this quantity:

$$
E_{q_1}\left[\frac{\tilde{p}(\boldsymbol{y},\theta_1)}{q_1(\theta_1)}\right] \le E_{q_1}\left[\frac{\tilde{p}(\boldsymbol{y},\theta_1)}{q_1(\theta_1)}\right] + \int q_1(\theta_1)\log\left(\frac{q_1(\theta_1)}{\tilde{p}(\theta_1|\boldsymbol{y})}\right)d\theta_1
$$

Observe that this upper bound is actually the logarithm of the marginal likelihood from the derivation in section 3.1. It should be obvious from the inequality that the quantity $\mathcal{F}[q]$ is maximized with respect to q_1 when this upper bound is attained. This occurs when:

$$\begin{aligned} q_1(\theta_1) = \tilde{p}(\theta_1|\boldsymbol{y}) &\propto \exp\left(\int \cdots \int q_2(\theta_2)...q_M(\theta_M) \log p(\boldsymbol{y}, \boldsymbol{\theta}) d\theta_2...d\theta_M\right) \\ &\propto \exp\left(E_{-\theta_1}[\log(\boldsymbol{y}, \boldsymbol{\theta})]\right) \end{aligned}$$

The proof for each of the other variational distributions is identical.

A.2 Expected Values for MFVB Algorithm

This section of the appendix gives the required expected values to fully define Algorithm 1 presented in Section 2.2. Complex expected value terms are approximated by applying a first order Taylor series approximation centered around the mean of the variable(s) of interest. For simplicity of notation, for any parameter or set of parameters θ, let $\theta^* = \mathrm{E}_\theta[\theta]$. Additionally, any expected value with respect to a covariate, $z_{i,t}$, apply for missing covariates. The following expected value approximations and calculations can be derived for an observed covariate, $z_{i,t}^{obs}$, by letting $\mathrm{E}[z_{i,t}] = z_{i,t}^* = z_{i,t}^{obs}$ and $\mathrm{Var}(z_{i,t}^{obs}) = 0$.

1. Expected Values Necessary for the Variational Density of d_i

The variational density of d_i involves two expected values that need to be approximated and one which can be computed directly:

$$\begin{aligned}
\mathrm{E}_{\beta_0,\beta_1,z_{i,t}}[\log(1-\phi_{i,t})] =&\mathrm{E}_{\beta_0,\beta_1,z_{i,t}}[\log(1-\mathrm{expit}(\beta_0+\beta_1 z_{i,t}))] \approx \log(\mathrm{expit}(\beta_0^*+\beta_1^* z_{i,t}^*))\\
\mathrm{E}_{\beta_0,\beta_1,z_{i,t}}[\log(\phi_{i,t})] =&\mathrm{E}_{\beta_0,\beta_1,z_{i,t}}[\log(\mathrm{expit}(\beta_0+\beta_1 z_{i,t}))] \approx \log(\mathrm{expit}(\beta_0^*+\beta_1^* z_{i,t}^*))\\
\mathrm{E}_{p_t}[\log(1-p_t)] =&\psi(\beta_{q^*(p_t)}) - \psi(\alpha_{q^*(p_t)}+\beta_{q^*(p_t)})
\end{aligned}$$

2. Expected Values Necessary for the Variational Density of p_t

The optimal variational density of p_t only requires computing $\mathrm{P}(t \leq d_i)$ for every $i \in 1,\dots,n$, which is simply a summation of cell probabilities from each d_i.

3. Expected Values Necessary for the Variational Density of Δ_t

The variational density of Δ_t only relies on simple expected values of parameters with normal and gamma variational distributions.

4. Expected Values Necessary for the Variational Density of τ

The variational density of τ involves computing the following expectation:

$$\mathrm{E}_{z,\Delta}[(z_{i,j}-z_{i,j-1}-\Delta_j)^2] =(z_{i,j}^*-z_{i,j-1}^*-\Delta_j^*)^2+\mathrm{Var}(z_{i,j})+\mathrm{Var}(z_{i,j-1})+\mathrm{Var}(\Delta_j)$$

5. Expected Values Necessary for the Variational Density of $\boldsymbol{\beta}$

The variational density of the coefficients for survival involve complex expectations with respect to a potentially missing covariate, particularly for computing the variance:

$$\begin{aligned}
\mathrm{E}_{z_{i,t}}[\log(\phi_{i,t})] =&\mathrm{E}_{z_{i,t}}[\log(\text{expit}(\beta_0 + \beta_1 z_{i,t}))] \\
\approx& \log(\text{expit}(\beta_0 + \beta_1 z^*_{i,t})) \\
\mathrm{E}_{z_{i,t}}[\log(1-\phi_{i,t})] =&\mathrm{E}_{z_{i,t}}[\log(\text{expit}(1-\beta_0 + \beta_1 z_{i,t}))] \\
\approx& \log(1-\text{expit}(\beta_0 + \beta_1 z^*_{i,t})) \\
\mathrm{E}_{z_{i,t}}[\text{expit}(\beta_0 + \beta_1 z_{i,t})&(1-\text{expit}(\beta_0 + \beta_1 z_{i,t}))] \\
&\approx \text{expit}(\beta_0 + \beta_1 z^*_{i,t})(1-\text{expit}(\beta_0 + \beta_1 z^*_{i,t})) \\
\mathrm{E}_{z_{i,t}}[z_{i,t}\text{expit}(\beta_0 + \beta_1 z_{i,t})&(1-\text{expit}(\beta_0 + \beta_1 z_{i,t}))] \\
&\approx z^*_{i,t}\text{expit}(\beta_0 + \beta_1 z^*_{i,t})(1-\text{expit}(\beta_0 + \beta_1 z^*_{i,t})) \\
\mathrm{E}_{z_{i,t}}[z^2_{i,t}\text{expit}(\beta_0 + \beta_1 z_{i,t})&(1-\text{expit}(\beta_0 + \beta_1 z_{i,t}))] \\
&\approx (z^*_{i,t})^2\text{expit}(\beta_0 + \beta_1 z^*_{i,t})(1-\text{expit}(\beta_0 + \beta_1 z^*_{i,t}))
\end{aligned}$$

Like the variational density for p_t, there are also expected values with respect to each d_i which are direct functions of cell probabilities.

6. Expected Values Necessary for the Variational Density of Missing Covariates

The variational densities for the missing covariates contain the following complicated expected values involving the coefficients on survival:

$$\begin{aligned}
\mathrm{E}_{\beta_0,\beta_1}[\log(\text{expit}(\beta_0 + \beta_1 z_{i,t}))] \approx& \log(\text{expit}(\beta^*_0 + \beta^*_1 z_{i,t})) \\
\mathrm{E}_{\beta_0,\beta_1}[\log(1-\text{expit}(\beta_0 + \beta_1 z_{i,t}))] \approx& \log(1-\text{expit}(\beta^*_0 + \beta^*_1 z_{i,t})) \\
\mathrm{E}_{\beta_0,\beta_1}[\beta_1^2\text{expit}(\beta_0 + \beta_1 z_{i,t})(1-\text{expit}(\beta_0 + \beta_1 z_{i,t}))] \approx&(\beta^*_1)^2\text{expit}(\beta^*_0 + \beta^*_1 z_{i,t}) \\
&\times (1-\text{expit}(\beta^*_0 + \beta^*_1 z_{i,t}))
\end{aligned}$$

Additionally, there are straightforward expected values of parameters with normal and gamma variational densities and also expected values with respect to each d_i which are direct functions of cell probabilities.

A.3 Derivation of Optimal Variational Densities for the MFVB Method Applied to the Mixed Effects Model

Beginning with the optimal variational density for $\mathbf{u}$ and $\boldsymbol{\beta}$, we have:

$$\begin{aligned}
q^*_{\boldsymbol{\beta},\mathbf{u}} \propto & \exp\Bigg(E_{\boldsymbol{\sigma}^2}\Bigg[-\frac{1}{2}\log\left((2\pi)^n(\sigma^2_\epsilon)^n\right) - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) \\
& \quad -\frac{1}{2}\log\left((2\pi)^{\sum K_l}\prod_{l=1}^{r}(\sigma^2_{ul})^{K_l}\right) - \frac{1}{2}\mathbf{u}^T\mathbf{G}^{-1}\mathbf{u} - \frac{1}{2}\boldsymbol{\beta}^T(\sigma^2_{\boldsymbol{\beta}}\mathbf{I}_p)^{-1}\boldsymbol{\beta}\Bigg]\Bigg) \\
q^*_{\boldsymbol{\beta},\mathbf{u}} \propto & \exp\Bigg(-\frac{1}{2}E_{\boldsymbol{\sigma}^2}\Bigg[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) + \mathbf{u}^T\mathbf{G}^{-1}\mathbf{u} + \boldsymbol{\beta}^T(\sigma^2_{\boldsymbol{\beta}}\mathbf{I}_p)^{-1}\boldsymbol{\beta}\Bigg]\Bigg) \\
q^*_{\boldsymbol{\beta},\mathbf{u}} \propto & \exp\Bigg(-\frac{1}{2}\Bigg((\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T\left(E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_\epsilon}\right]\mathbf{I}_n\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) \\
& \quad + \mathbf{u}^T\text{blockdiag}\left(E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{u1}}\right]\mathbf{I}_{K_1},\ldots,E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{ur}}\right]\mathbf{I}_{K_r}\right)\mathbf{u} + \boldsymbol{\beta}^T(\sigma^2_{\boldsymbol{\beta}}\mathbf{I}_p)^{-1}\boldsymbol{\beta}\Bigg)\Bigg) \\
q^*_{\boldsymbol{\beta},\mathbf{u}} \propto & \exp\Bigg(-\frac{1}{2}\Bigg((\boldsymbol{\beta},\mathbf{u})^T\text{blockdiag}\left(\frac{1}{\sigma^2_{\boldsymbol{\beta}}}\mathbf{I}_p, E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{u1}}\right]\mathbf{I}_{K_1},\ldots,E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{ur}}\right]\mathbf{I}_{K_r}\right)(\boldsymbol{\beta},\mathbf{u}) \\
& \quad (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T\left(E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_\epsilon}\right]\mathbf{I}_n\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})\Bigg)\Bigg) \\
q^*_{\boldsymbol{\beta},\mathbf{u}} \propto & \exp\Bigg(-\frac{1}{2}\Bigg((\boldsymbol{\beta},\mathbf{u})^T\text{blockdiag}\left(\frac{1}{\sigma^2_{\boldsymbol{\beta}}}\mathbf{I}_p, E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{u1}}\right]\mathbf{I}_{K_1},\ldots,E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{ur}}\right]\mathbf{I}_{K_r}\right)(\boldsymbol{\beta},\mathbf{u}) \\
& \quad E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_\epsilon}\right](\mathbf{y}-\mathbf{C}(\boldsymbol{\beta},\mathbf{u}))^T(\mathbf{y}-\mathbf{C}(\boldsymbol{\beta},\mathbf{u}))\Bigg)\Bigg)
\end{aligned}$$

where $\mathbf{C} = [\mathbf{X},\mathbf{Z}]$. Completing the square yields:

$q^*_{\boldsymbol{\beta},\mathbf{u}}$ is $\mathcal{N}(\mu_{q_{\boldsymbol{\beta},\mathbf{u}}},\Sigma_{q_{\boldsymbol{\beta},\mathbf{u}}})$

where

$$\Sigma_{q_{\boldsymbol{\beta},\mathbf{u}}} = \left(\text{blockdiag}\left(\frac{1}{\sigma^2_{\boldsymbol{\beta}}}\mathbf{I}_p, E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{u1}}\right]\mathbf{I}_{K_1},\ldots,E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_{ur}}\right]\mathbf{I}_{K_r}\right) + E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_\epsilon}\right]\mathbf{C}^T\mathbf{C}\right)^{-1}$$

and

$$\mu_{q_{\boldsymbol{\beta},\mathbf{u}}} = \Sigma_{q_{\boldsymbol{\beta},\mathbf{u}}}\left(E_{\boldsymbol{\sigma}^2}\left[\frac{1}{\sigma^2_\epsilon}\right]\mathbf{C}^T\mathbf{y}\right).$$

The optimal variational density for $\boldsymbol{\sigma}^2$ is derived by:

$$
\begin{aligned}
q^*_{\boldsymbol{\sigma^2}} \propto \exp\Bigg(& E_{\boldsymbol{\beta},\mathbf{u}}\Bigg[-\frac{n}{2}\log(\sigma^2_\epsilon) - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu}) \\
& -\frac{1}{2}\log\left(\prod_{l=1}^{r}(\sigma^2_{ul})^{K_l}\right) - \frac{1}{2}\mathbf{u}^T\mathbf{G}^{-1}\mathbf{u} + \sum_{l=1}^{r}\left[(-A_{ul}-1)\log(\sigma^2_{ul}) - \frac{B_{ul}}{\sigma^2_{ul}}\right] \\
& + (-A_\epsilon - 1)\log(\sigma^2_\epsilon) - \frac{B_\epsilon}{\sigma^2_\epsilon}\Bigg]\Bigg) \\
q^*_{\boldsymbol{\sigma^2}} \propto \exp\Bigg(& -\frac{n}{2}\log(\sigma^2_\epsilon) - \frac{1}{2}E_{\boldsymbol{\beta},\mathbf{u}}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})\right] \\
& -\frac{1}{2}\log\left(\prod_{l=1}^{r}(\sigma^2_{ul})^{K_l}\right) - \frac{1}{2}E_{\boldsymbol{\beta},\mathbf{u}}[\mathbf{u}^T\mathbf{G}^{-1}\mathbf{u}] + \sum_{l=1}^{r}\left[(-A_{ul}-1)\log(\sigma^2_{ul}) - \frac{B_{ul}}{\sigma^2_{ul}}\right] \\
& + (-A_\epsilon - 1)\log(\sigma^2_\epsilon) - \frac{B_\epsilon}{\sigma^2_\epsilon}\Bigg) \\
q^*_{\boldsymbol{\sigma^2}} \propto \exp\Bigg(& -\frac{n}{2}\log(\sigma^2_\epsilon) - \frac{1}{2}\left(\mathrm{tr}(\mathbf{R}^{-1}\mathbf{C}\Sigma_{q(\beta,u)}\mathbf{C}^T) + (\mathbf{y}-\mathbf{C}\boldsymbol{\mu}_{q(\beta,u)})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{C}\boldsymbol{\mu}_{q(\beta,u)})\right) \\
& -\frac{1}{2}\log\left(\prod_{l=1}^{r}(\sigma^2_{ul})^{K_l}\right) - \frac{1}{2}\left(\mathrm{tr}(\mathbf{G}^{-1}\Sigma_{q(u)}) + \boldsymbol{\mu}^T_{q(u)}\mathbf{G}^{-1}\boldsymbol{\mu}_{q(u)}\right) \\
& + \sum_{l=1}^{r}\left[(-A_{ul}-1)\log(\sigma^2_{ul}) - \frac{B_{ul}}{\sigma^2_{ul}}\right] + (-A_\epsilon - 1)\log(\sigma^2_\epsilon) - \frac{B_\epsilon}{\sigma^2_\epsilon}\Bigg) \\
q^*_{\boldsymbol{\sigma^2}} \propto \exp\Bigg(& \left(-\frac{n}{2} - A_\epsilon - 1\right)\log(\sigma^2_\epsilon) + \sum_{l=1}^{r}\left(-\frac{K_l}{2} - A_{ul} - 1\right)\log(\sigma^2_{ul}) \\
& - \frac{1}{\sigma^2_\epsilon}\left(\frac{1}{2}\mathrm{tr}(\mathbf{C}\Sigma_{q(\beta,u)}\mathbf{C}^T) + \frac{1}{2}(\mathbf{y}-\mathbf{C}\boldsymbol{\mu}_{q(\beta,u)})^T(\mathbf{y}-\mathbf{C}\boldsymbol{\mu}_{q(\beta,u)}) + B_\epsilon\right) \\
& - \sum_{l=1}^{r}\frac{1}{\sigma^2_{ul}}\left(\frac{1}{2}\mathrm{tr}(\Sigma_{q(ul)}) + \frac{1}{2}\boldsymbol{\mu}^T_{q(ul)}\boldsymbol{\mu}_{q(ul)} + B_{ul}\right)
\end{aligned}
$$

I now have the kernel of a product of inverse gamma distributions and can determine that:

$$
\begin{aligned}
& q^*_{\sigma^2_\epsilon} \text{ is IG}\left(\frac{n}{2} + A_\epsilon, \frac{1}{2}\mathrm{tr}(\mathbf{C}\Sigma_{q(\beta,u)}\mathbf{C}^T) + \frac{1}{2}(\mathbf{y}-\mathbf{C}\boldsymbol{\mu}_{q(\beta,u)})^T(\mathbf{y}-\mathbf{C}\boldsymbol{\mu}_{q(\beta,u)}) + B_\epsilon\right) \\
& q^*_{\sigma^2_{ul}} \text{ is IG}\left(\frac{K_l}{2} + A_{ul}, \frac{1}{2}\mathrm{tr}(\Sigma_{q(ul)}) + \frac{1}{2}\boldsymbol{\mu}^T_{q(ul)}\boldsymbol{\mu}_{q(ul)} + B_{ul}\right).
\end{aligned}
$$

This completes the derivations of the optimal variational densities when applying the MFVB method to the mixed effects model.

A.4 Temporary Emigration

In Section 4.5, I found that when fitting the truncated CJS model to the cliff swallows data, smaller values of k yielded estimates of survival parameters that were systematically lower and capture parameters that were systematically higher than those obtained when fitting the full, non-truncated CJS model. As I did not see this effect in any of the simulation scenarios presented in Section 4.4, this suggests that the cliff swallows data violates an assumption necessary to fit CJS-type models. One of these assumptions, detailed in Section 1.3, is that all death and emigration from the population is permanent. In this section, I explore the implications of fitting the truncated CJS model in the presence of temporary emigration by simulating data that allows individuals, who are alive, to freely enter and exit the population.

I simulated data sets containing $n = 3000$ individuals observed over $T = 34$ capture occasions. The simulated data follow a CJS model with capture probabilities, $p_2, \ldots, p_{34}$, and survival probabilities, $\phi_1, \ldots, \phi_{33}$, that vary by capture occasion, with one important deviation: individuals may leave and re-enter the catchable population after they are first captured. Specifically, I consider random emigration and immigration such that individuals who are alive and members of the study population may exit the population on any capture occasion with probability η and individuals who are alive and have exited the population may re-enter the population on any subsequent capture occasion with probability ν. For simplicity, I also assume that individuals inside and outside of the population have a common probability of survival, ϕ_t.

The true capture probabilities were drawn from a uniform distribution ranging from 0.3 to 0.7 while the true survival probabilities were drawn from a uniform distribution ranging from 0.5 to 0.9. After determining the capture and survival probabilities, 3 different data sets were generated: one with $\eta = 0.4$ and $\nu = 0.2$, one with $\eta = 0.3$ and $\nu = 0.3$, and one with $\eta = 0$ (i.e. no temporary emigration). I then fit the truncated CJS model to each data set with truncation parameters of $k = 3$, $k = 5$, $k = 10$, and $k = T$ (no truncation).

The results for the analysis of the data set with no temporary emigration, seen in Figure A.1 for capture probabilities and Figure A.2 for survival probabilities, are consistent with what I observed in my earlier simulation study (described in Section 4.4). There appears to be no systematic bias in the capture or survival probability estimates with respect to k. Although, setting $k = 3$ does seem to result in parameter estimates that are more variable than the estimates produced by fitting the non-truncated model and truncated models with k=5 or k=10. You can also observe, by looking at the relationship of the parameter estimates to the true parameter values (indicated by the gray horizontal lines), that no value of k results in parameter estimates consistently closer to the true parameter values than the others.

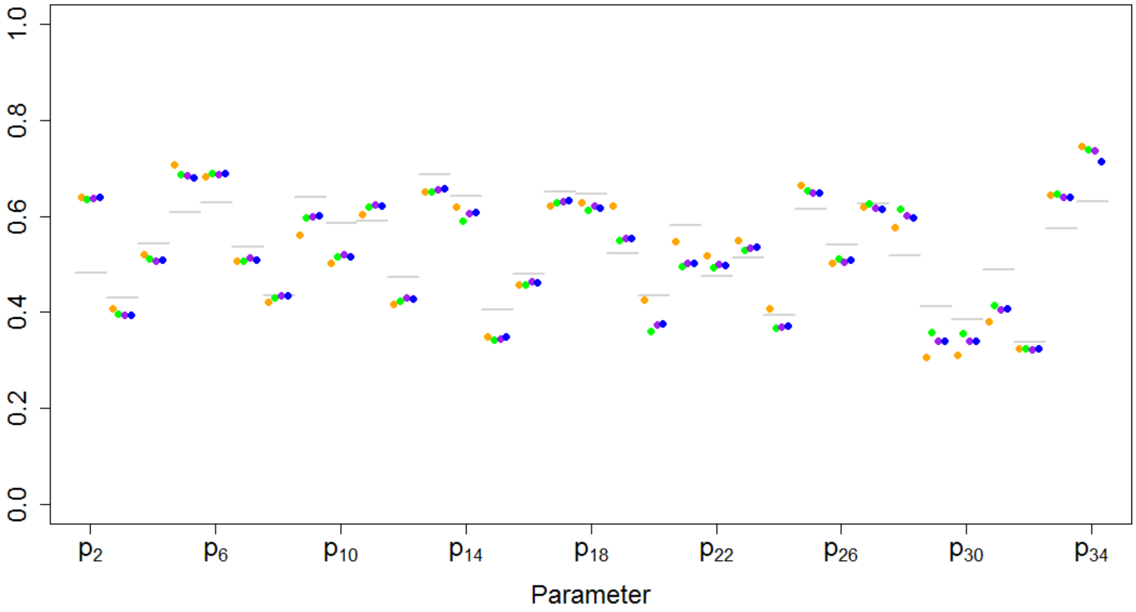


Figure A.1: *Estimates of Capture Parameters from the Truncated CJS Model for a Single Simulated Data Set with No Temporary Emigration*

Parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the capture parameters of a simulated data set with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was no temporary emigration present in this simulated data set. The parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

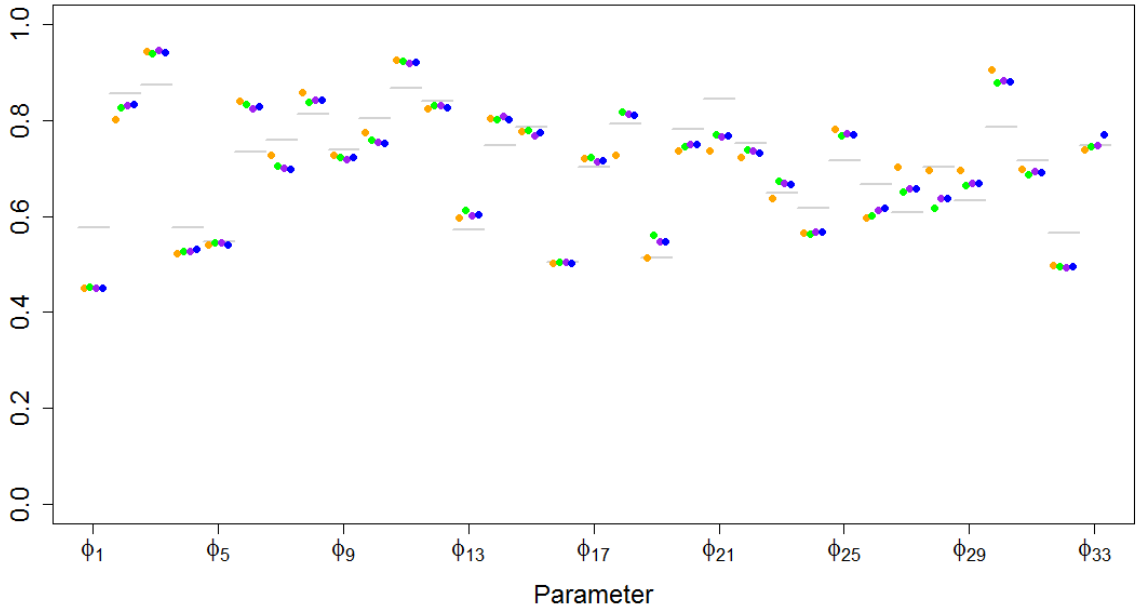


Figure A.2: *Estimates of Survival Parameters from the Truncated CJS Model for a Single Simulated Data Set with No Temporary Emigration*

Parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the survival parameters of a simulated data set with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was no temporary emigration present in this simulated data set. The parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

This is not the case, however, once temporary emigration is introduced into the simulated data sets. Figures A.3 and A.4 display the estimated capture and survival probabilities obtained when fitting the truncated and non-truncated CJS models to data in which individuals may leave the population with probability $\eta = 0.4$ and re-enter the population with probability $\nu = 0.2$. I now see that smaller values of k (i.e. more extreme truncation of capture histories) lead to consistently higher estimates of capture probability and consistently lower estimates of survival probability. Not only is this trend consistent across all capture occasions, but more extreme truncation also appears to produce less accurate estimates of survival probability. Capture probabilities are consistently underestimated by all of the models in the presence of temporary emigration, as one would expect. This result can also be observed in Figures A.5 and A.6, which display estimated capture and survival probabilities obtained when fitting the models to data where individuals have a leave probability of $\eta = 0.3$ and re-entry probability of $\nu = 0.3$.

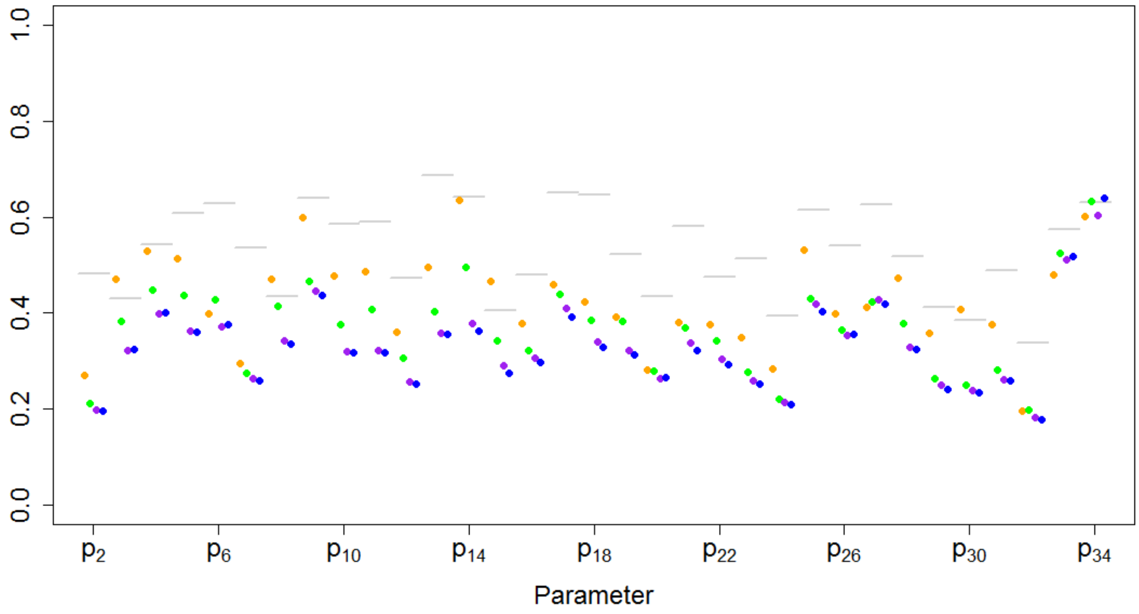


Figure A.3: *Estimates of Capture Parameters from the Truncated CJS Model for a Single Simulated Data Set with Severe Temporary Emigration*

Parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the capture parameters of a simulated data set with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in this simulated data set, with individuals leaving the population with probability $\eta = 0.4$ and re-entering the population with probability $\nu = 0.2$. The parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

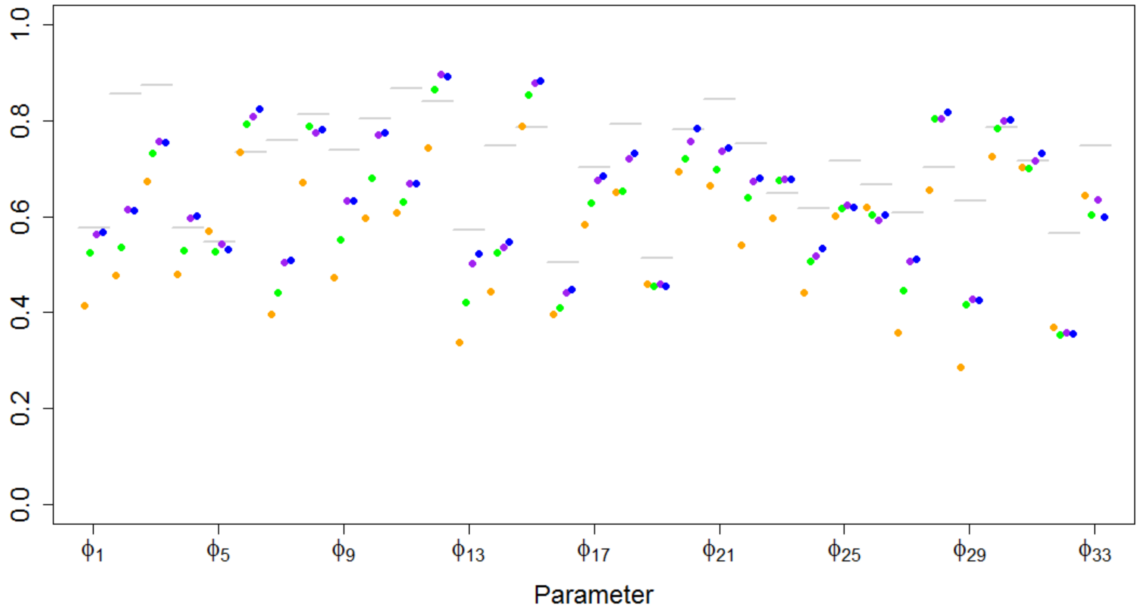


Figure A.4: *Estimates of Survival Parameters from the Truncated CJS Model for a Single Simulated Data Set with Severe Temporary Emigration*

Parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the survival parameters of a simulated data set with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in this simulated data set, with individuals leaving the population with probability $\eta = 0.4$ and re-entering the population with probability $\nu = 0.2$. The parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

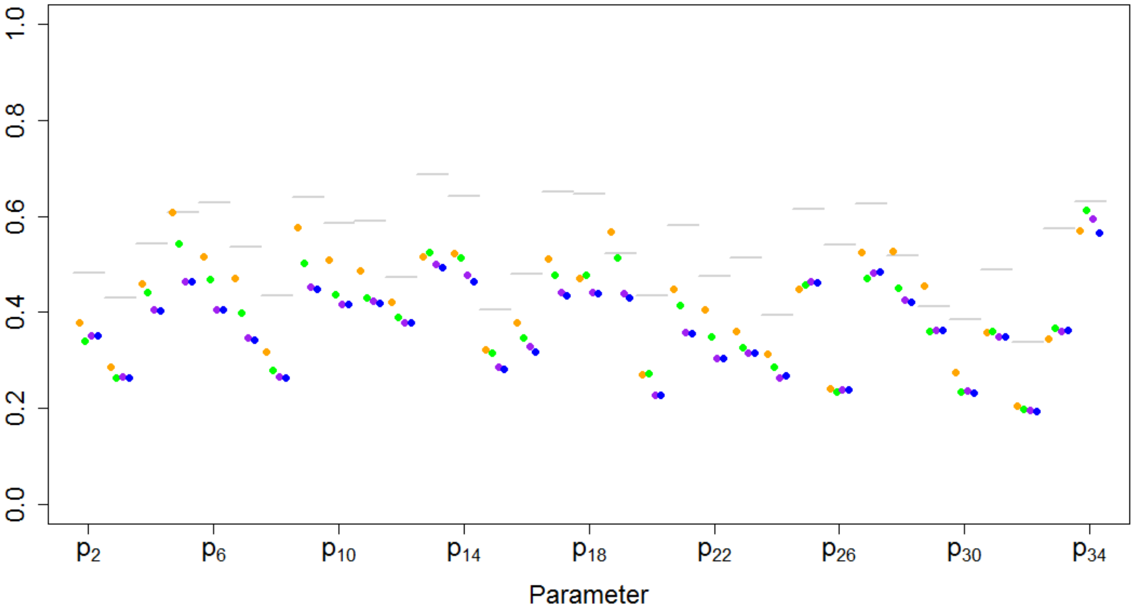


Figure A.5: *Estimates of Capture Parameters from the Truncated CJS Model for a Single Simulated Data Set with Moderate Temporary Emigration*

Parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the capture parameters of a simulated data set with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in this simulated data set, with individuals leaving the population with probability $\eta = 0.3$ and re-entering the population with probability $\nu = 0.3$. The parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

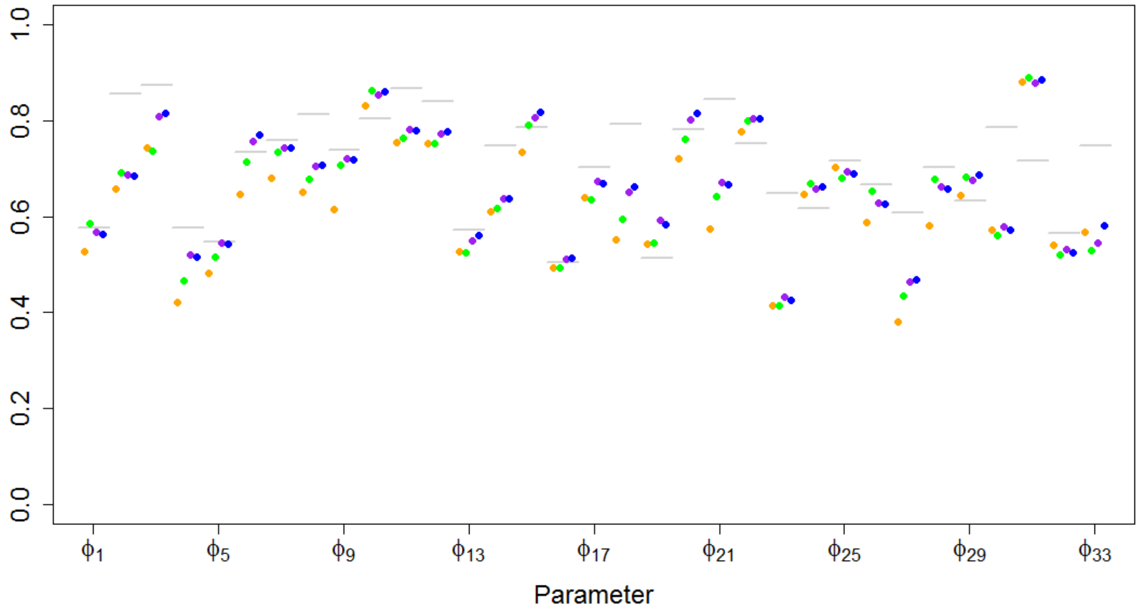


Figure A.6: *Estimates of Survival Parameters from the Truncated CJS Model for a Single Simulated Data Set with Moderate Temporary Emigration*

Parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the survival parameters of a simulated data set with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in this simulated data set, with individuals leaving the population with probability $\eta = 0.3$ and re-entering the population with probability $\nu = 0.3$. The parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

Additionally, to further illustrate these conclusions, I performed simulations from the aforementioned three different temporary emigration scenarios on a larger scale, simulating from each scenario 100 times. Figures A.7 and A.8 display the estimates of capture and survival probability, respectively, averaged over 100 simulated data sets without any temporary emigration present. Figures A.9 and A.10, however, display the averaged parameter estimates from data sets where temporary emigration was present, with individuals leaving the population with probability $\eta = 0.4$ and re-entering the population with probability $\nu = 0.2$ Lastly, Figures A.11 and A.12 show the corresponding plots when individuals have a leave probability of $\eta = 0.3$ and re-entry probability of $\nu = 0.3$. Note that these are the same three scenarios depicted in the plots introduced before, but the parameter estimates are averaged across 100 simulated data sets, reducing variability and making the effects of different values of k more clear.

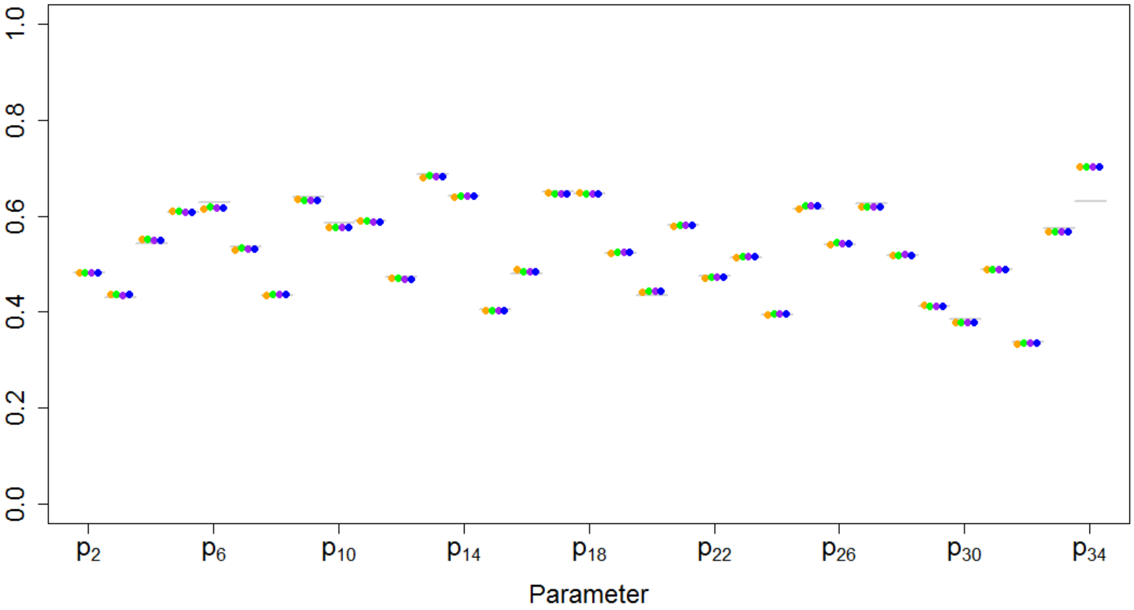


Figure A.7: *Average Estimates of Capture Parameters from the Truncated CJS Model across 100 Simulated Data Sets with No Temporary Emigration*

Average parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the capture parameters from 100 simulated data sets with $n = 3000$ *individuals observed over* $T = 34$ *capture occasions. There was no temporary emigration present in these simulated data sets. The average parameter estimates from fitting the truncated CJS model with* $k = 3$ *(orange),* $k = 5$ *(green),* $k = 10$ *(purple), and the non-truncated CJS model (*$k = T$*, blue) are presented. Parameter labels are located on the x-axis.*

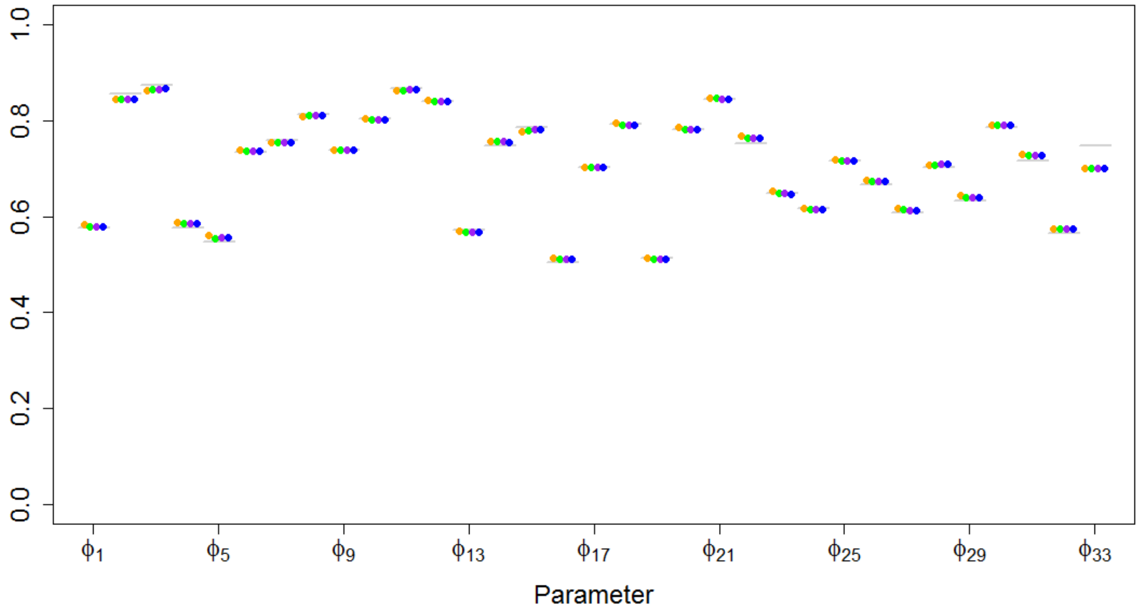


Figure A.8: *Average Estimates of Survival Parameters from the Truncated CJS Model across 100 Simulated Data Sets with No Temporary Emigration*

Average parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the survival parameters from 100 simulated data sets with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was no temporary emigration present in these simulated data sets. The average parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

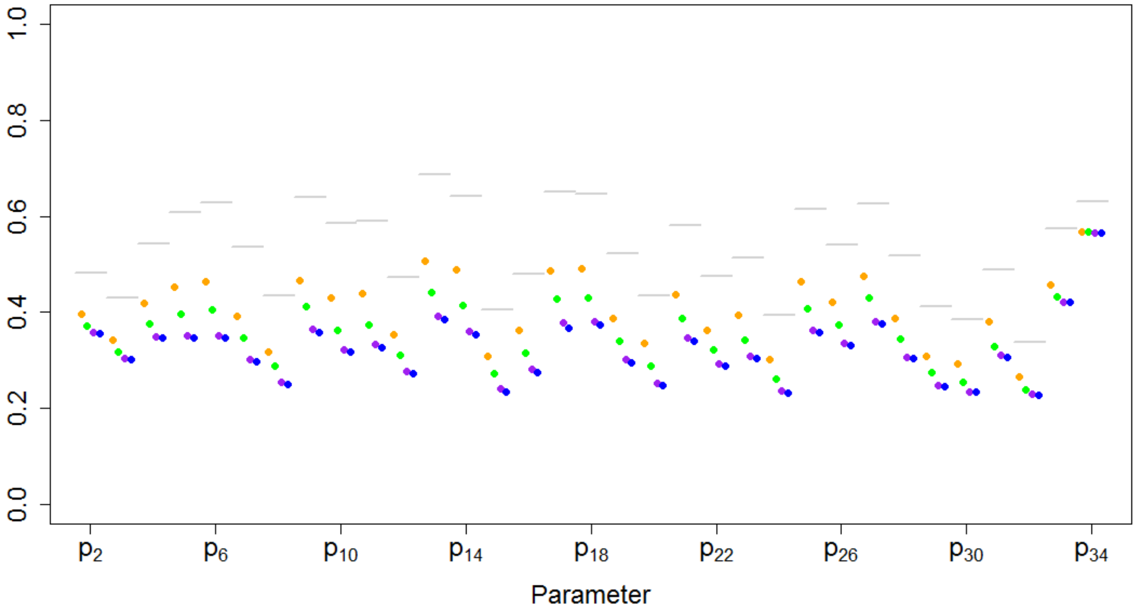


Figure A.9: *Average Estimates of Capture Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Severe Temporary Emigration*

Average parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the capture parameters from 100 simulated data sets with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in these simulated data sets, with individuals leaving the population with probability $\eta = 0.4$ and re-entering the population with probability $\nu = 0.2$. The average parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

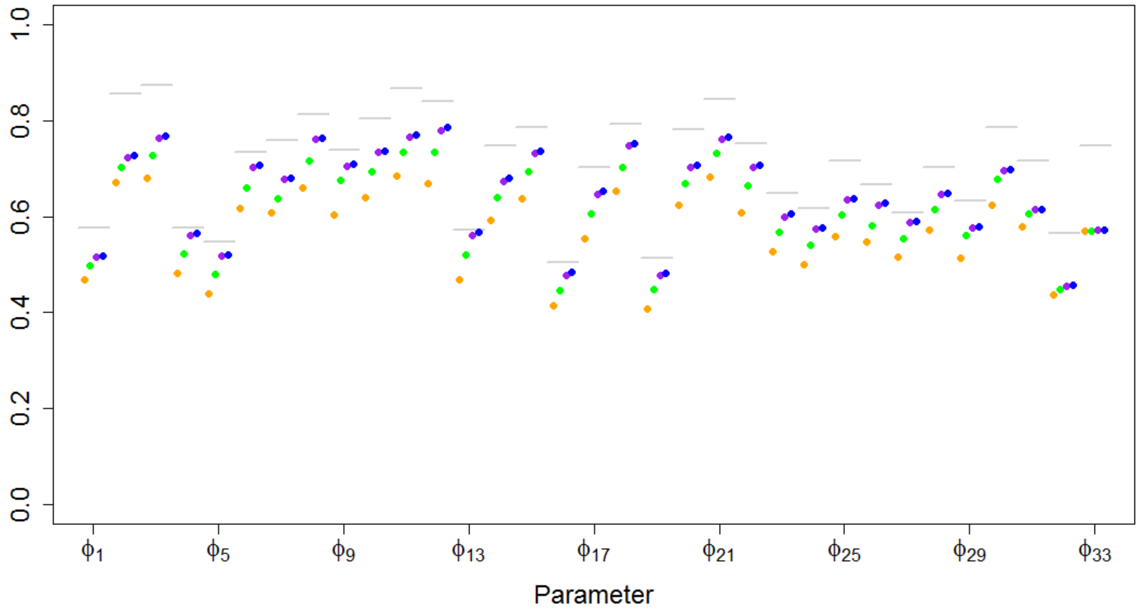


Figure A.10: *Average Estimates of Survival Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Severe Temporary Emigration*

Average parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the survival parameters from 100 simulated data sets with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in these simulated data sets, with individuals leaving the population with probability $\eta = 0.4$ and re-entering the population with probability $\nu = 0.2$. The average parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

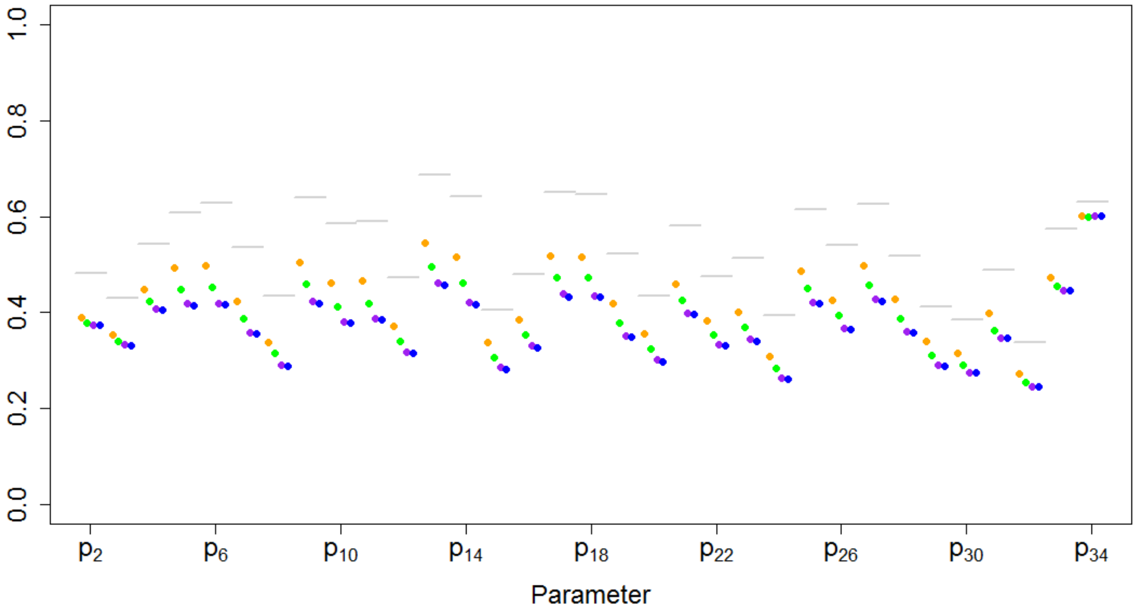


Figure A.11: *Average Estimates of Capture Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Moderate Temporary Emigration*

Average parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the capture parameters from 100 simulated data sets with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in these simulated data sets, with individuals leaving the population with probability $\eta = 0.3$ and re-entering the population with probability $\nu = 0.3$. The average parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

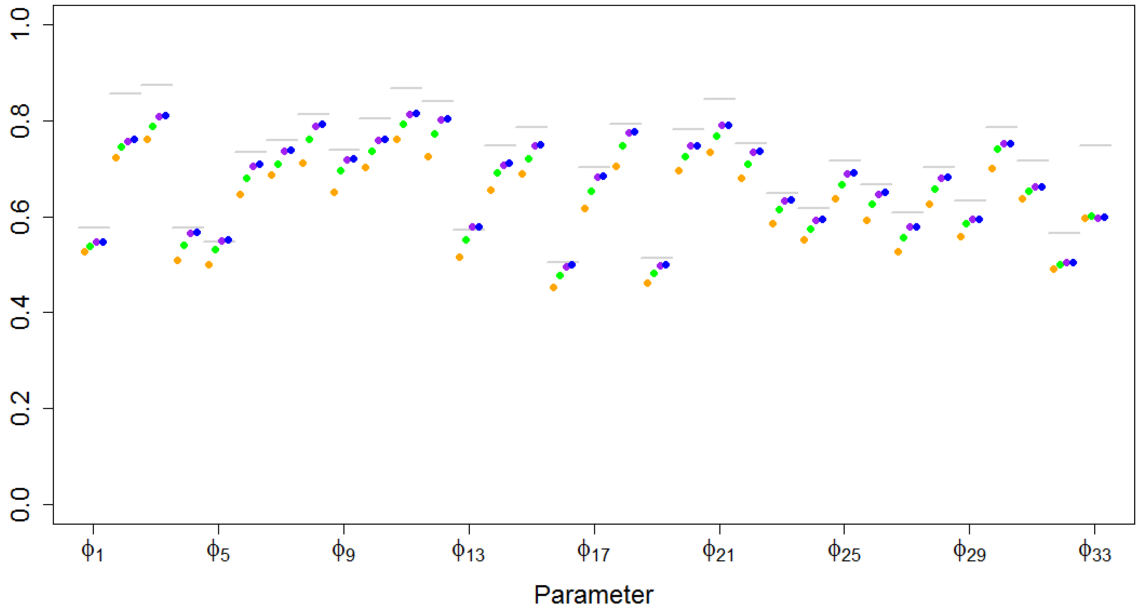


Figure A.12: *Average Estimates of Survival Parameters from the Truncated CJS Model across 100 Simulated Data Sets with Moderate Temporary Emigration*

Average parameter estimates (colored points) and true underlying parameter values (horizontal gray lines) for the survival parameters from 100 simulated data sets with $n = 3000$ individuals observed over $T = 34$ capture occasions. There was temporary emigration present in these simulated data sets, with individuals leaving the population with probability $\eta = 0.3$ and re-entering the population with probability $\nu = 0.3$. The average parameter estimates from fitting the truncated CJS model with $k = 3$ (orange), $k = 5$ (green), $k = 10$ (purple), and the non-truncated CJS model ($k = T$, blue) are presented. Parameter labels are located on the x-axis.

The bias observed in the estimated capture and survival probabilities obtained when fitting the truncated CJS model at low values of k makes intuitive sense. If an individual leaves the population only to eventually return, a truncated capture history is less likely to contain both that individuals exit and re-entry to the study population than a full, non-truncated CJS model or a truncated CJS model with a higher value of k. Therefore, the truncated capture history would be assigned a higher probability that the individual had died when that individual had truly only left the population. The same capture history observed in the full, non-truncated CJS model, however, may eventually see that individual again, after the individual has returned to the population, and instead conclude that the individual was simply not captured, and estimate artificially low capture probabilities instead. This result is also consistent with the systematic bias I observed when examining the cliff swallows data. This indicates that the cliff swallows data may exhibit temporary emigration of individuals from the study population and, if this is the case, than the survival parameters estimated when fitting the truncated CJS model are likely to be more affected by this temporary emigration at lower values of k.

Bibliography

Simon J Bonner. Continuous, individual, time-dependent covariates in the cormack-jolly-seber model. Master's thesis, Simon Fraser University, 2003.

Simon J Bonner, Byron JT Morgan, and Ruth King. Continuous covariates in mark-recapture-recovery analysis: a comparison of methods. *Biometrics*, 66(4):1256–1265, 2010.

SJ Bonner and CJ Schwarz. An extension of the Cormack–Jolly–Seber model for continuous covariates with application to *Microtus pennsylvanicus*. *Biometrics*, 62 (1):142–149, 2006.

Charles R Brown and Mary Bomberger Brown. *Coloniality in the cliff swallow: the effect of group size on social behavior*. University of Chicago Press, 1996.

C Brownie, JE Hines, JD Nichols, KH Pollock, and JB Hestbeck. Capture-recapture studies for multiple strata including non-markovian transitions. *Biometrics*, pages 1173–1187, 1993.

Stephen T Buckland. Monte carlo confidence intervals. *Biometrics*, pages 811–817, 1984.

Stephen T Buckland and Paul H Garthwaite. Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, pages 255–268, 1991.

Kenneth P Burnham. *Design and analysis methods for fish survival experiments based on release-recapture*. American Fisheries Society, 1987.

K.P. Burnham and D.R. Anderson. *Model Selection and Inference: A Practical Information-theoretic Approach*. Intelligence, SS. of Lncs; 1501. Springer, 1998. ISBN 9780387985046. URL `https://books.google.com/books?id=OHHwAAAAMAAJ`.

Edward A Catchpole, Byron JT Morgan, and Giacomo Tavecchia. A new method for analysing discrete life history data with missing covariate values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):445–460, 2008.

Tim H Clutton-Brock and Josephine M Pemberton. *Soay sheep: dynamics and selection in an island population*. Cambridge University Press, 2004.

RM Cormack. Estimates of survival from the sighting of marked animals. *Biometrika*, pages 429–438, 1964.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

Samuel Gershman, Matt Hoffman, and David Blei. Nonparametric variational inference. *arXiv preprint arXiv:1206.4665*, 2012.

Justin Grimmer. An introduction to bayesian inference via variational approximations. *Political Analysis*, page mpq027, 2010.

Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

George M Jolly. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, pages 225–247, 1965.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37 (2):183–233, 1999.

R King, SP Brooks, and T Coulson. Analyzing complex capture–recapture data in the presence of individual and temporal covariates and model uncertainty. *Biometrics*, 64(4):1187–1195, 2008.

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic variational inference in stan. In *Advances in neural information processing systems*, pages 568–576, 2015.

Roland Langrock, Ruth King, et al. Maximum likelihood estimation of mark–recapture–recovery models in the presence of continuous covariates. *The Annals of Applied Statistics*, 7(3):1709–1732, 2013.

Jean-Dominique Lebreton, Kenneth P Burnham, Jean Clobert, and David R Anderson. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological monographs*, 62(1):67–118, 1992.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

Stephan Mandt and David Blei. Smoothed gradients for stochastic variational inference. In *Advances in Neural Information Processing Systems*, pages 2438–2446, 2014.

C.A. McGrory, D.M. Titterington, R. Reeves, and A.N. Pettitt. Variational Bayes for estimating the parameters of a hidden potts model. *Statistics and Computing*, 19(3):329–340, 2009.

Thomas P Minka. Using lower bounds to approximate integrals. *Informal notes available at http://www. stat. cmu. edu/˜ minka/papers/learning. html*, 2001.

James D Nichols, John R Sauer, Kenneth H Pollock, and Jay B Hestbeck. Estimating transition probabilities for stage-based population projection matrices using capture-recapture data. *Ecology*, 73(1):306–312, 1992.

JT Ormerod and MP Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

Carl Georg Johannes Petersen. The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station*, 6:1–48, 1896.

Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2014. URL `http://CRAN.R-project.org/package=nlme`. R package version 3.1-117.

Martyn Plummer. Jags: A program for analysis of Bayesian graphical models using Gibbs sampling, 2003.

Güngör Polatkan, Mingyuan Zhou, Lawrence Carin, David Blei, and Ingrid Daubechies. A bayesian nonparametric approach to image super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):346–358, 2015.

Kenneth H Pollock. The use of auxiliary variables in capture-recapture modelling: an overview. *Journal of Applied Statistics*, 29(1-4):85–102, 2002.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

Carl J Schwarz, Jake F Schweigert, and A Neil Arnason. Estimating migration rates using tag-recovery data. *Biometrics*, pages 177–193, 1993.

George AF Seber. A note on the multiple-recapture census. *Biometrika*, pages 249–259, 1965.

George AF Seber. *The Estimation of Animal Abundance*. The Blackburn Press, 2002.

Chong Wang and David M Blei. Variational inference in nonconjugate models. *The Journal of Machine Learning Research*, 14(1):1005–1031, 2013.

Ian R White and John B Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29 (28):2920–2931, 2010.

John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

Hannah Worthington, Ruth King, and Stephen T Buckland. Analysing mark–recapture–recovery data in the presence of missing covariate data via multiple imputation. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(1): 28–46, 2015.

Liqun Xi, Ray Watson, Ji-Ping Wang, and Paul SF Yip. Estimation in capture–recapture models when covariates are subject to measurement errors and missing data. *Canadian Journal of Statistics*, 37(4):645–658, 2009.

Vita

Woodrow W. Burchett

Education

University of Kentucky (Lexington, KY)
  M.S. in Statistics, 2012
Georgetown College (Georgetown, KY)
  B.S. in Mathematics, 2010

Employment

Research Statistician, 2016 - Present
  Pfizer, Groton, Connecticut
Research Assistant in the Applied Statistics Lab, 2012-2016
  Department of Statistics, University of Kentucky
Summer Intern, 2013
  Kentucky Energy and Environment Cabinet, Frankfort, Kentucky
Teaching Assistant, 2010-2012
  Department of Statistics, University of Kentucky