Continuous, Individual, Time-Dependent Covariates in the Cormack-Jolly-Seber Model

by

Simon J. Bonner

B.Sc. Hons., McGill University, 2001

A Project submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
April 2003

APPROVAL

Name: Simon J. Bonner

Degree: Master of Science

Title of Project: Continuous, Individual, Time-Dependent Covariates in the Cormack-Jolly-Seber Model

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Carl Schwarz
Senior Supervisor
Simon Fraser University

Dr. K. Laurence Weldon
Simon Fraser University

Dr. Tim Swartz
External Examiner
Simon Fraser University

Date Approved: _____

Abstract

Capture-recapture methods are common in ecological studies when it is impossible to completely enumerate a population. Instead, individuals are captured, marked, and released back into the environment. Information about the population is derived by comparing the ratios of tagged to untagged individuals on subsequent capture occasions. Of primary interest in most capture-recapture studies are parameters governing the survival rate.

Recent developments have focused on techniques to study the impact of different covariates on the survival probabilities. Straightforward methods have been developed for cases where the covariates are either discrete, constant over time, or apply to the population as a whole, but the problem has not been solved for the case of continuous time-dependent covariates unique to each individual. Because it is impossible to measure such variables on occasions where an individual was not captured, estimation can be considered a missing data problem. In this project, a model was constructed using a diffusion process to describe changes in a univariate covariate. Logistic functions were used to link the covariate to capture and survival rates and incorporate the data into the Cormack-Jolly-Seber model. Two methods of parameter estimation were developed based on techniques commonly used to handle missing data: a frequentist approach using the EM-algorithm and a Bayesian approach through MCMC. These methods were applied to simulated and real data-sets, and comparison of the results is provided. In short, though both methods yield similar results in most cases, the Bayesian approach was more reliable, easier to implement, and yielded estimates of the standard errors more readily than the EM-algorithm.

Acknowledgments

I want first to recognize the work of my supervisor, Dr. Carl Schwarz. Carl deserves the credit for the idea behind this project and for much of the insight that powered my work. Throughout the project Carl complimented the work I had done and continued to push me forward. I always left his office with more things to do but even more conviction to do them.

My examining committee members, Dr. Tim Swartz and Dr. K. Larry Weldon, gave useful comments and criticisms of my written project. Dr. Swartz also assisted me in producing the Bayesian estimation scheme for the model I developed.

The sample data was provided by Dr. Evan Cooch of Cornell University. Dr. Cooch also offered guidance on the analysis and answered all of my questions about the habits of snow geese living at La Pérouse Bay.

I have enjoyed my time at SFU, and for that I must thank my friends and fellow students. You all made work fun and distracted me when it was time for a rest. Jacquie Gregory has listened to my ramblings, even when they didn't make any sense, and has been the best alternate dance partner. Michael Lo and I struggled together to complete our projects at the same time, and I have felt him watching out for me for all my time here. They are both great friends. Special thanks to May Lee, Wilson and Helen Lu, Monica and Rene Norena, and Jason Nielsen. Acknowledgements also to all the members of The Improbabilities. I looked forward to our volleyball games each week, and second place was far above my expectations.

I am very grateful for the support of all those closest to me. Though they were far away in distance, my parents, Ruth and Ron, knew when I needed to talk. Thank you for the phone calls, the e-mails, and the letters with their little gifts. My sister,

Sarah, was also always willing to listen and to cheer me on. Thank you. My girlfriend, Alyssa MacLean, was the first to have to deal with my highs and my lows. She saw me through the crises and kept me going when everything seemed backward and words came in the wrong order. I understand now how difficult writing can be. Alyssa, thank you for moving so we could be together for the past 8 months. It would have been much harder without you here.

One final thank you again to Carl Schwarz, who originally got me interested in studying at SFU. Carl brought me out to visit and told me that this would be a great place both to learn and to have fun. He was right.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The goal of ecological research is to understand the demographic properties of a population by estimating parameters including the population size and birth and death rates. In most situations, especially those involving large animals or large study areas, a complete enumeration of the population is impossible and sampling methods must be used. A popular alternative is to employ a capture-recapture study which involves capturing samples of individuals, marking them, releasing them back into the wild, and making inferences based on the captures of marked and unmarked individuals over the following periods. Many different variations of the basic capture-recapture experiment have been developed to improve parameter estimates and to handle cases where the assumptions of simple capture-recapture models cannot be satisfied. Several researchers have focused their studies on incorporating extra information from different variables into capture-recapture methods. This project introduces methods for studying heterogeneity in capture and survival rates caused by dependence on dynamic individual continuous covariates.

1.1 Basic Capture-Recapture and the Cormack-Jolly-Seber Model

The most basic capture-recapture study is one that includes only two capture events and identical markers for all individuals. A sample of animals from the population is trapped on the first occasion, marked, and then released back into the study area. After allowing time for the tagged animals to disperse, a second sample is collected that includes both marked and unmarked individuals. Assuming that the population is closed to births, deaths, and movement the size of the population will not change between the two capture occasions and is estimated by equating the ratio of tagged to untagged individuals in the second sample with the ratio in the general population. This leads to the Lincoln-Petersen estimator, first used by Laplace in 1878 and formalized by F. C. Lincoln for studying North-American water-fowl populations circa 1930 (Williams et al., 2002).

Studying survival rates requires more complicated capture-recapture designs with $\tau > 2$ capture occasions. Researchers must also be able to distinguish individuals, using unique markers or individual characteristics, in order to generate the capture histories for each individual, vector's of 1's and 0's that indicate the capture occasions when each animal was observed. In total, there are 2^τ possible capture histories. For a study involving 5 capture occasions, one possible capture history is 01010 indicating that this animal was first captured and marked on the 2^{nd} occasion, recaptured on the 4^{th} occasion, and not observed at times 1, 3, or 5. This example highlights some of the difficulties that arise in studying open populations (i.e. those that change due to birth, death, and animal movement). At the 1^{st} and 5^{th} capture occasions it is impossible to know if this animal was alive or simply went unobserved. On the 3^{rd} occasion one can conclude that the animal was alive, but it is still not possible to know if the animal was inside the study area. The capture history for individual i is denoted by the vector ω_i and has t^{th} element ω_{it}.

The dominant model used for studying survival based on capture histories is the Cormack-Jolly-Seber (CJS) model, developed simultaneously by 3 researchers in the mid-1960's (Cormack, 1964; Jolly, 1965; Seber, 1965). To handle the uncertainties

regarding the status of individuals before their first capture, the CJS model conditions on each individual's first release and considers only the events on subsequent occasions. By nature, this ignores all individuals who were never captured and about whom nothing is known. It also ignores the processes of recruitment in the population and concentrates only on animal losses. Ambiguities of the type that occur on the 3^{rd} occasion of the example are avoided by assuming that all emigration out of the study area is permanent. To describe the events of each life history, Cormack, Jolly, and Seber introduce two sets of probability parameters common to every individual in the population. Capture probabilities, p_t, model the probability that an animal alive in the study area at time t is captured, and survival probabilities, ϕ_t, model the probability that an individual alive at time t is still alive at time $t+1$. For the above history, the likelihood contribution to the CJS model is:

$$L_{01010} = \phi_2(1-p_3) \cdot \phi_3 p_4 \cdot ((1-\phi_4) + \phi_4(1-p_5))$$

The first two terms in this expression describe the events that were known to occur on times 3 and 4, and the final term accounts for both the possibility that the individual died between times 4 and 5 and that it survived but was not observed. No information is included for either times 1 or 2 because of the conditioning on the release at the 2^{nd} occasion.

Even with conditioning on the first release times for each individual, it is still not possible to distinguish between animals that die between capture times, those that leave the study area and do not return, and those that remain alive and in the study area but simply evade capture. This means that the exact time of death will never be known for any individual. Because of this confounding effect between death and emigration, Williams et al. (2002) suggests that the parameter ϕ_t be called the "apparent survival" as it does not reflect the true survival rate of individuals in the population. Other confounding effects mean that some parameters in the model are not identifiable. For example, without further capture occasions it is not possible to estimate the number of animals that remain in the population at capture time τ, and the final parameters for capture and survival, p_τ and ϕ_{t-1}, cannot be distinguished. In total, the full CJS model with τ capture occasions involves $2\tau -$

3 estimable parameters. Traditionally these have been estimated using maximum likelihood techniques (Lebreton et al., 1992) though Bayesian methods have also been developed (Poole, 2002).

Implicit in the CJS model are several more assumptions regarding the population. Williams et al. (2002) list 6 including the assumptions that individual capture events are independent between individuals and capture times, marks are never lost or misidentified, and changes in the population occur only between the capture occasions. Throughout the 1980's and early 1990's different researchers have published papers that introduced modifications of the capture-recapture design or the CJS model to account for situations that do not satisfy these assumptions. One area of specific interest is the development of models that account for differences in the capture and survival probabilities across the population. A common approach is to allow the probabilities to depend on different covariates measured from the captured individual or the environment.

1.2 Previous Models using Covariate Data

In capture-recapture experiments field researchers frequently measure variables that might be associated with the survival or capture probability of the animals in the population. In studies of migrating birds, for example, it is common practise for banders to record a captured bird's sex, body mass, wing length, fat quantity, and other pertinent information describing the bird's health. Environmental variables like temperature or daily precipitation and even the bird's location may also be associated with the likelihood that a bird alive at one occasion is captured or survives until the next occasion. In their overview of capture-recapture methods, Williams et al. (2002) initially classify possible covariates into 4 categories by sample space (discrete or continuous) and by time dependence (static or dynamic). For dynamic variables they later add the distinction between those which are deterministic and those which are stochastic in nature. A full treatment must also distinguish between variables that are unique to the individual, like body mass, and those that are common to the entire population, like most environmental measurements. Except for the case

of continuous, dynamic, individual covariates to be discussed in this project, statisticians and ecologists have already developed different models to incorporate all of this information into capture-recapture analysis.

Surprisingly, little attention was given to incorporating covariates in early capture-recapture development. One of the first papers to consider using covariate data in models of closed populations was written by Pollock et al. (1984). In this paper, Pollock and Nichols consider several types of auxiliary variable including continuous environmental variables, static individual covariates, and deterministic individual covariates. These are the simplest variables to work with because they can be determined at every time point for the whole population or for each individual provided that the individual is captured at least once. Information from the variables is incorporated into the model using a logistic link function:

$$log\left(\frac{p}{1-p}\right) = \beta' X$$

where p is the probability of capture, X is the vector of covariates, and β the coefficients. In most cases, the parameters of the logistic model and the corresponding capture rates can be estimated using maximum likelihood analysis. Survival rates need not be modelled because the population is assumed to be closed.

One difficulty that arises in the model of Pollock et al. (1984) for individual covariates is that no information is available for the individuals in the population that were never captured. In the paper, Pollock and Nichols use a full likelihood approach that bases inference on the histories of all individuals in the closed population. For the simplest case where the coefficients of the logistic function are assumed constant over time, the likelihood in terms of the total population size, N, is proportional to:

$$L = \prod_{i=1}^{N} \prod_{t=1}^{\tau} p(x_{it})^{\omega_{it}} (1 - p(x_{it}))^{\omega_{it}}$$

In this expression x_{it} represents the vector of the individual covariates for subject i and the environmental covariates at time t. Many of these values are unknown because the individuals are never observed, and if any of the covariates are continuous the full likelihood cannot be used without modification. To overcome this problem Pollock

et al. (1984) suggest an admittedly "ad hoc" method (Pollock, 2002) that requires continuous variables to be discretized, but a more appropriate solution was proposed by Huggins (1989). Similar to the development of the CJS model, Huggins uses a model that conditions on the first capture event for each individual and the new conditional likelihood becomes:

$$L = \prod_{i=1}^{n} \prod_{t=1}^{\tau} p(x_{it})^{\omega_{it}} (1 - p(x_{it}))^{\omega_{it}}$$

where $i = 1, \ldots, n$ indexes only those individuals captured during the study. Once more, Huggins employs a logistic model to incorporate the covariates into the capture probability and provides MLE's for the different parameters. The one drawback of conditioning in this manner is that the value N no longer appears in the likelihood and cannot be estimated by likelihood methods. Instead, the population size is estimated by the method of moments (Pollock, 2002).

Another solution to the problem of missing individual covariates is provided by Borchers et al. (1998). Borchers considers the problem of analyzing line transect surveys where the number of individuals in the study area is constant, but the probability of observing each animal depends upon some covariates, most obviously the distance from the animal to the transect line. While line transect surveys do not exactly match capture-recapture theory, they are remarkably similar to closed population experiments and the methods developed by Borchers do relate. In order to use the full likelihood including all individuals in the population, Borchers assumes that the covariates are independently and identically distributed according to some distribution function, $\pi(\cdot)$, specified up to a set of unknown parameters. Borchers computes the probability of being observed at least once in the survey, $p_\cdot = \int p(x)\pi(x)dx$, and then constructs the likelihood using a multinomial distribution over the possible capture histories where all unobserved individuals fall into a single category to themselves (Pollock, 2002):

$$L = \left[\frac{N!}{n!(N-n)!} \right] p_\cdot^{\,n} (1 - p_\cdot)^{(N-n)} \prod_{i=1}^{n} \prod_{t=1}^{\tau} \frac{p(x_{it})^{\omega_{it}} (1 - p(x_{it})^{\omega_{it}}) \pi(x_{it})}{p_\cdot}$$

As before, the capture probability is formed as a logistic function of the covariates. Modelling the distribution of the covariates does create a more complicated model, of

course, because the extra parameters that determine the exact distribution must be estimated along with the coefficients of the logistic regression. In some cases this can lead to very difficult integration problems when several covariates are used, and some models may have unidentifiable parameters (Pollock, 2002). However, this method does allows Borchers to use a model incorporating information from all individuals in the closed population and to compute the sample size using maximum likelihood techniques.

The conditioning method proposed by Huggins (1989) bears much resemblance to the CJS model and is easily extended for use with open populations. Lebreton et al. (1992) describe an adaptation of the CJS model that uses external variables to adjust both survival and capture through link functions. While the paper stresses the logistic model, it also suggests other possibilities like the log link, $f(x) = log(x)$, and identity link, $f(x) = x$, that are commonly used in generalized linear models. The advantage of the logistic function in capture-recapture is that the estimated rates of survival and capture always remain between 0 and 1, regardless of the covariate and coefficient values (Lebreton et al., 1992). As with Huggins's model, Lebreton defines the likelihood conditional upon the first capture for each individual and uses it to produce MLE's for the model parameters. The paper also describes more complex models with multiple covariates often using dummy variables to introduce categorical variables and of interaction terms to study how the different groups of individuals to react to the external variables. Formal tests are also described that can be used to chose between different models and determine which sets of variables have significant effect on the survival and capture rates.

More difficult to incorporate into the CJS model are time-dependent covariates which are determined stochastically for each individual. Even when these variables are discrete they can be difficult to handle and conditioning on the captured individuals is no longer sufficient to provide a likelihood from which estimates can be computed. Not only are some values of the covariates unknown because individuals remain unobserved during the entire study, but they are also unknown on each capture time that a previously marked individual is not recaptured. To compensate for this it is necessary to employ a more complex model like that of Borchers et al. (1998) to describe the

distribution of the missing covariate values.

For discrete covariates of this nature the distribution can be modelled using a Markov chain describing the probability of moving from one state to any other. Interestingly, this method was first used in modelling open populations by Arnason (1973) long before the work on closed populations by Borchers et al. (1998). Arnason's method was later modified by Schwarz et al. (1993), and is now known as the Arnason-Schwarz model (Pollock, 2002). As motivation, Schwarz considers a population living in a study area that is broken into K different areas between which the animals can migrate. In each area the individuals are assumed to have the same capture probability though it may vary between capture events. The probability of survival from time t to $t+1$ is assumed to depend only on the animal's location at time t as is the probability of moving from location k to l in the same period (Williams et al., 2002). These three quantities are denoted by p_t^k,ϕ_t^k, and ψ_t^{kl} respectively. In a 3 stage study, an animal that is captured in location k on the first occasion and then recaptured on the third occasion in location l, would give the conditional likelihood contribution:

$$L_i = \phi_1^k \left[\sum_{r=1}^{K} \psi_1^{kr} (1 - p_2^r) \psi_2^{rl} \phi_2^r \right] p_3^l$$

Here the likelihood contribution is complicated by the sum that accounts for all moves the individual could have made between the 1^{st} and 3^{rd} capture events. Combining the contributions from all captured animals, Schwarz et al. (1993) defines the new conditional likelihood and derives expressions for the MLE's of the survival, capture, and transition parameters.

Since its development, different researchers have used extensions of the Arnason-Schwarz model to account for different situations. Brownie et al. (1993) introduce what they term a "memory" model that uses a two-stage Markov chain to describe the movements between locations. Of course, any number of previous locations could be used, but the number of transition parameters grows geometrically with the number of stages and would quickly exhaust the information in the data set. Several studies have also used the Arnason-Schwarz model in an attempt to incorporate continuous time-dependent individual covariates into the CJS model. An example of this is

Nichols et al. (1992) study of the meadow vole, *Microtus pennsylvanicus*. In this study, body mass is used as a covariate by breaking the measurements into 4 discrete classes assuming the effect on capture and survival is the same in each class, and Nichols does find a significant difference between the classes. This does not represent a true method for incorporating continuous covariates because information is bound to be lost when categorizing a continuous variable (Pollock, 2002).

In the same paper, Nichols also introduces an alternative method for handling multi-state covariates in CJS model. This method is based on the robust capture-recapture design which combines the techniques for both open and closed populations (Williams et al., 2002). In the experiment, Nichols arranges the capture occasions so that several events occur very close together, followed by a larger break, and then several more very close capture events. The population is assumed to be closed over the tightly grouped capture occasions, known as secondary periods, and open over the large time gaps, called primary periods. Using data from adjacent secondary periods in a single primary period, Nichols estimates the size of the population at each location, the proportion of individuals that have migrated from each of the other locations, and the capture rate for these animals. Transition and survival parameters are then estimated by comparing these numbers across the different primary periods (Nichols et al., 1992).

One problem that often arises in estimating the parameters for these CJS based models is that the likelihood expressions can become very complicated making the estimates very difficult to compute. As an example, consider an Arnason-Schwarz model with many different states and individuals who remain hidden for long periods of time. This would result in a likelihood expression with many missing values requiring many sums to account for all possible unobserved transitions. A way to simplify this is to consider the model as a missing data problem and work instead with the complete data likelihood. In the case of the Arnason-Schwarz model, the complete data set constitutes all observed data as well as the unobserved locations for each of the individuals and knowledge of whether the animal is dead or alive for each occasion after it is first captured. A common method that takes advantage of the missing data approach is the expectation-maximization (EM) algorithm originally developed as a

synthesis of existing methods in the late 1970's (Dempster et al., 1977) The algorithm uses a two step iterative procedure to solve for the MLE's and often greatly simplifies the computations by maximizing a derived likelihood that is usually much simpler in form (see section 2.3 for more details). Application of the EM-algorithm to capture-recapture experiments is explored by Van Deusen (2002). Van Deusen focuses specifically on the model used by Lebreton et al. (1992) to incorporate covariate data into the CJS model using logistic regression, but in theory the same method could be applied to any of the extensions discussed.

Several researchers in the area of capture-recapture methodology have also developed Bayesian methods of parameter estimation in place of the frequentist maximum likelihood approach. A Bayesian method for estimating the population size of a closed population as an extension of the Lincoln-Peterson estimator was first published in 1972. For the more complex models described here, the effort to develop Bayesian methods has been led by French statistician Jerome Dupuis. In 1995 Dupuis described a Bayesian method for estimating the parameters in the Arnason-Schwarz model that uses the technique of Gibbs sampling to generate a sample from the posterior distribution of the parameters (Dupuis, 1995). Efforts in this area continue today.

With all of these extensions and estimation techniques, the category of continuous, individual, time dependent covariates has still been overlooked. Several methods report to include this type of variable, but further inspection shows that they always resort to some form of simplification. The classic approaches in the literature are either to categorize these variables into discrete classes as in Nichols et al. (1992) or to assume that the variables are constant over time (see the application of Piper (2002) as an example). In either case, some information is lost and this might mask the true relationship between the variable and the survival and capture probabilities. The following chapter describes a new model based on the techniques described here to include these variables in the CJS model without having to force them into a simpler form.

Chapter 2

Methodology

2.1 Data Structure and Notation

For simplicity I will assume that the discrete capture times are evenly spaced, denoting them $1, \ldots, \tau$, and that a total of n individuals are captured on at least one occasion. The variables t and i will be used throughout to index the capture time and individual respectively. When discussing a single individual as an example the subscript i will often be dropped to simplify notation.

All data gathered for a single individual, i, is recorded in three vectors of length τ. As for the basic CJS, the capture history, denoted ω_i, is a vector of indicator variables such that the tth entry, ω_{it}, equals 1 if the individual was captured at time t and 0 otherwise. The survival history vector, $\mathbf{s_i}$, records the occasions when the individual was known to be alive in the study area. If the individual is known to be alive at time t then $\mathbf{s_{it}} = 1$, and otherwise the missing data marker, $\cdot$, is used indicating that the survival status is unknown. Similarly, the vector $\mathbf{z_i}$ is used to record the information about the covariates and has entry z_{it}, possibly vector valued itself, when the individual was captured at time t and $\cdot$ otherwise. Together, these three vectors contain all information regarding the capture, survival, and covariate measurements for the i^{th} individual.

As an example of this notation, consider the capture history discussed in section

1.1. For this individual the 3 data vectors are:

$$\omega_i = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad s_i = \begin{pmatrix} \cdot \\ 1 \\ 1 \\ 1 \\ \cdot \end{pmatrix}, \quad \text{and} \quad z_i = \begin{pmatrix} \cdot \\ z_{i2} \\ \cdot \\ z_{i4} \\ \cdot \end{pmatrix}$$

To aid with notation, I also introduce two summary statistics to index the first and last time an individual is captured: $a_i = \min\{t : \omega_{it} = 1\}$ and $b_i = \max\{t : \omega_{it} = 1\}$. For this individual $a_i = 2$ and $b_i = 4$. Further, parameters relating to the model of the covariates, survival, and capture will be introduced as they arise in the following sections.

2.2 CJS Drift Model

The model I will use to include continuous, time-dependent, individual covariates in the CJS model builds on many of the concepts discussed in section 1.2. As with the case of discrete time-dependent individual covariates it is necessary to model the covariate distribution to account for the unobserved values. To define this distribution, I develop a drift process similar to Brownian motion that leads to a Markov chain with continuous state space, the logical extension of the Arnason-Schwarz model. Following the example of (Lebreton et al., 1992) and (Huggins, 1989) the capture and survival rates are linked to the covariate values through logistic regression. This type of model is suggested as future work by Pollock (2002), but specifics of how to model the change in the covariate over time and how to produce estimates are not considered.

To keep things simple I will work only in the univariate case, though the results should easily extend for use with multiple continuous covariates. As motivation consider the obvious examples of a time-dependent, individual, continuous covariate: body mass. Imagine also a population of individuals all of whom have reached maturity and all living in the same area subject to the same external forces. In good years

with plentiful food and little competition all animals gain mass and in poor years all animals lose mass. Animals in the population are expected to react in a similar manner, though some variation about the mean is bound to occur due to individual differences.

To describe the dynamics of this process of gain and loss I use a model based on the Weiner process. This process is an extension of the simple random walk to continuous time and was first used to describe the random movement of particles. Cox and Miller (1965) define the Weiner process as a stochastic process, $X(t)$ for $0 \leq t \leq \tau$, which satisfies two properties:

i) for any $t_1 < t_2$, $X(t_2) - X(t_1) \sim N\left(\mu(t_2 - t_1), \sigma^2(t_2 - t_1)\right)$

ii) for any $t_1 < t_2 < t_3 < t_4$, $X(t_2) - X(t_1)$ and $X(t_4) - x(t_3)$ are independent

The variable μ in these expressions is called the drift and determines the general trend in the process. If $\mu = 0$ then the process will oscillate randomly about 0 whereas if $\mu > 0$ the process will, on average, result in increasing values of $X(t)$ as time progresses and vice-versa for $\mu < 0$. How similarly different instances of the chain behave is determined by the variance, σ^2. In application to capture-recapture experiments one can imagine that body mass is constantly fluctuating and this process runs continuously for all animals. The observations on the capture times form discrete snapshots for each individual defining a new process $Z_t = X(t)$ for the discrete times $t = 1, \ldots, \tau$. From property (i) it follows that the distribution of the values in this process are determined by the differences between subsequent capture occasions. Under the assumption that all intervals have unit length $Z_t - Z_{t-1}$ is normally distributed with mean μ and variance σ^2. A special case of property (ii) is that for each $t = 1, \ldots, \tau$ the difference $Z_t - Z_{t-1}$ is independent of the differences for all previous intervals. Writing $Z_t = Z_{t-1} + (Z_t - Z_{t-1})$ it follows that $Z_t | Z_{t-1}, \ldots, Z_0 \sim Z_t | Z_{t-1}$ and hence the process forms a Markov chain. To allow for different conditions between each pair of capture occasions I introduce separate drift parameters, $\mu_1, \ldots, \mu_{\tau-1}$, such that:

$$Z_t | Z_{t-1} = z_{t-1} \sim N(z_{t-1} + \mu_{t-1}, \sigma^2)$$

From this distribution the conditional likelihood contribution for an individual captured once and then on all remaining occasions is simply a product of univariate normal density functions.

The model becomes more complicated when we consider the possibility that an individual evades capture on some occasions. First, consider the situation for an individual which is last captured at time $b < \tau$. For each $\Delta = 0, \ldots, \tau - b$ the conditional distribution of $Z_{b+\Delta}|Z_b = z_b$ is defined by writing $Z_{b+\Delta}$ as a sum of independent normal random variables:

$$Z_{b+\Delta} = z_b + \sum_{r=b}^{\Delta-1}(Z_{r+1} - Z_r)$$

As expected, this shows that $Z_{b+\Delta} \sim N(z_b + \sum_{r=b}^{\Delta-1} \mu_r, \Delta\sigma^2)$. In order to build the joint density of the unobserved values it is necessary to consider the correlation between the variables. From 2.2 it follows that the random vector $\mathbf{Z}^* = (Z_{b+1}, \ldots, Z_\tau)^T$ can be written in terms of linear combinations of independent $N(0,1)$ random variables, $W_1, \ldots, W_{\tau-b}$:

$$\mathbf{Z}^* = \begin{pmatrix} Z_b + \mu_b \\ Z_b + \mu_b + \mu_{b+1} \\ \vdots \\ Z_b + \sum_{r=b}^{\tau-1} \mu_r \end{pmatrix} + \sigma \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_{\tau-b} \end{pmatrix} \tag{2.1}$$

By definition the distribution of $\mathbf{W}^*$ given Z_b is multivariate normal with mean:

$$E(\mathbf{Z}|Z_b = z_b) = \left(z_b + \mu_b, z_b + \mu_b + \mu_{b+1}, \ldots, z_b + \sum_{r=b}^{\tau-1} \mu_r\right)^T$$

and variance-covariance matrix:

$$Var(\mathbf{Z}^*|Z_b = z_b) = \sigma^2 \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & 3 & \cdots & \tau - b \end{bmatrix}$$

When several missing observations occur in the middle of the history it becomes necessary to condition not only on the previous observation but also on the observation directly following. Suppose an individual is observed at times t and $t+s+1$ but not on the s occasions between. Using the same argument as in equation 2.1 the distribution of $\mathbf{Z}^{**} = (Z_{i,t+1}, \ldots, Z_{i,t+s})^T$ conditional on both Z_{it} and $Z_{i,t+s+1}$ is again multivariate normal with mean:

$$E(\mathbf{Z}^{**}|Z_t = z_t, Z_{t+s+1} = z_{t+s+1}) = \frac{1}{s+1}\begin{pmatrix} (s)(z_t + \mu_t) + (z_{t+s+1} - \sum_{r=t+1}^{t+s} \mu_r) \\ (s-1)(z_t + \mu_t + \mu_{t+1}) + 2(z_{t+s+1} - \sum_{r=t+2}^{t+s} \mu_r) \\ \vdots \\ (Z_t + \sum_{r=t}^{t+s-1} \mu_r) + s(z_{t+s+1} - \mu_{t+s}) \end{pmatrix}$$

and variance-covariance matrix:

$$Var(\mathbf{Z}^{**}|Z_t = z_t, Z_{t+s+1} = z_{t+s+1}) = \frac{\sigma^2}{s+1}\begin{bmatrix} s & s-1 & s-2 & \ldots & 1 \\ s-1 & 2(s-2) & 2(s-3) & \ldots & 2 \\ s-2 & 2(s-2) & 3(s-3) & \ldots & 3 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 2 & 3 & \ldots & s \end{bmatrix}.$$

In general, I will refer to the probability density function of a set missing covariate values conditional on observed information simply as $f(\cdot|\cdot)$. In most cases it is clear from context which random variables are being considered and only when necessary for clarity will they be included as subscripts.

As in the methods described in section 1.2 involving simpler covariate types, modelling the capture and survival follows a GLM strategy. As in (Lebreton et al., 1992) I use a logistic model such that given the covariate value for individual i at time t, z_{it}, the survival and capture probabilities are:

$$\begin{aligned} \phi_{it} &= \phi(z_{it}) = \frac{\exp(\beta_{s0} + \beta_{s1} z_{it})}{1 + \exp(\beta_{s0} + \beta_{s1} z_{it})} \\ p_{it} &= p(z_{it}) = \frac{\exp(\beta_{c0} + \beta_{c1} z_{it})}{1 + \exp(\beta_{c0} + \beta_{c1} z_{it})} \end{aligned}$$

In the GLM framework it is common to write these functions as inverses:

$$\begin{aligned}\mu_s &= \beta_{s0} + \beta_{s1} z_{it} = \log\left(\frac{\phi_{it}}{1-\phi_{it}}\right) \\ \mu_c &= \beta_{c0} + \beta_{c1} z_{it} = \log\left(\frac{p_{it}}{1-p_{it}}\right)\end{aligned}$$

where μ_s and μ_c are the linear predictors of survival and capture. The parameters β_s and β_c are of primary interest since they describe how the individual behaviour varies with the covariate.

Modelling in this manner assumes that both survival from t to $t+1$ and capture at time t are dependent only upon the value of the covariate at time t. This is highly plausible for the capture events which are assumed to be instantaneous, but it seems likely that survival depends at least on all values of the covariate between the two capture times. However, a model that incorporates the exact path is impossible in capture-recapture because individuals can only be observed on the discrete capture occasions. Instead, I view the quantity ϕ_{it} as an expected survival probability accounting for all paths starting at the value z_{it} at time t. Apart from the assumption of homogeneity of the capture and survival rates the other assumptions of the basic CJS model are maintained. In particular, I assume that capture occasions are instantaneous, marks are never lost or mistakes made in recording data, trap deaths do not occur, and individual histories are independent of one another.

Estimation of the parameters in this model using maximum likelihood methods similar to those used for the basic CJS model would require manipulating the likelihood contribution for each individual conditional on first capture times. Analogous to the Arnason-Schwarz model, the conditional likelihood is formed by considering all possible transitions between the observed weights to replace the missing data. Instead of summing over discrete classes the continuous covariate case requires integration over the continuous state space. Using the example capture history $\omega_i = (0,1,0,1,0)$ again, the likelihood contribution conditional on the first capture is:

$$\begin{aligned}L_i &= \phi(z_{i2}) \cdot \int_{-\infty}^{\infty} (1-p(z_3)) f(z_3|z_{i2}) \phi(z_3) f(z_{i4}|z_3) dz_3 \\ &\quad \cdot p(z_{i4}) \cdot \left((1-\phi(z_{i4})) + \int_{-\infty}^{\infty} (1-p(z_5)) f(z_5|z_{i4}) dz_5\right)\end{aligned} \tag{2.2}$$

The functional forms of ϕ_{it} and p_{it} have been used here to emphasize the dependence of the different parameters on the unknown values. For this individual alone it can be seen that maximization is not straightforward, even though in this simple case the multivariate distribution is avoided because the missing observations occur singly. In cases where an animal has not been captured on several subsequent occasions all paths must be considered giving rise to multiple integrals of arbitrary dimension. For this problem, maximizing the likelihood using standard techniques becomes impossible. Instead, I turn first to the EM-algorithm and follow the approach of Van Deusen (2002), and then I consider a Bayesian approach to the problem.

2.3 Estimation via EM-algorithm

One method used by many statisticians for estimation in complex missing data problems is the Expectation-Maximization (EM) algorithm, a frequentist technique that provides an iterative solution to finding the MLE's. The EM was first formalized by Dempster et al. (1977) in which the authors develop the algorithm as a generalization of several existing methods and demonstrate it's use in a variety of problems. They also provide theory showing that estimates from the EM-algorithm are guaranteed to converge to the MLE's as the number of iterations increases, provided that a few, weak regularity conditions are satisfied.

The basic problem considered by Dempster et al. (1977) is estimation of a set of parameters, θ, in situations where the observed data is in some way incomplete. The general structure of the problem involves a set of observed data, $\mathbf{y}$, and a set of unobserved data, $\mathbf{z}$, whose combination forms the complete data-set, $\mathbf{x} = \mathbf{y} \cup \mathbf{z}$. It is assumed that a one-to-many mapping exists between the observed and complete sample spaces such that that the observed data does not uniquely define the complete data set. In the simplest form, this includes problems where data from a set of individuals is missing at random and can take many different true values, exactly as in the capture-recapture situation. More complex situations, like problems involving latent variables which are inherently missing for all observations but whose knowledge simplifies the data structure, also fall under the general framework.

Rather than working with the likelihood based on the observed data, inference from the the EM-algorithm relies on the complete data log-likelihood (CDLL). Letting $p(x;\theta)$ be the density of the complete data, the CDLL is simply $l = \log(p(x;\theta))$ considered as a function of the parameter values. Of course, based on the observed data it is impossible to know the true value of the CDLL for any set of parameters. To overcome this the EM-algorithm iterates between two steps; conceptually these involve constructing an estimate to the CDLL based on the current parameter estimates and then updating these estimates by maximization. In the first step, called the Expectation or E-step, the value of the CDLL is approximated by computing its expected value conditional on the observed data and the current parameter estimates. On the kth iteration with the previous parameter estimates denoted by $\theta^{(k-1)}$ this yields a function of the parameter values:

$$Q(\theta|\theta^{(k-1)}) = E(\log p(\mathbf{x};\theta)|\mathbf{y},\theta).$$

In the Maximization or M-step, new estimates are computed by treating $Q(\theta|\theta^{(k-1)})$ as if it were the true CDLL and maximizing to find the MLE's. Iteration provides a sequence of estimates for the parameters, $\theta^{(k)}$, which converge to the true MLE values as $k \to \infty$. In practise, the algorithm is iterated many times until some convergence criterion is met and the final parameter values are used as estimates.

The obvious advantage of the EM-algorithm in the capture-recapture problem incorporating continuous covariates is the substitution of the CDLL for the complex observed likelihood. In this scenario completing the data set involves two steps: filling in the missing covariate values and determining the survival status. Since all estimation is performed conditional on the first capture occasion, this is only done for $t > a_i$ for each individual. Combining the notation developed above and in section 2.1, I will denote the complete data-set involving full records of the capture, survival, and covariate status for all individuals by $\mathbf{x}$, and the full set of parameters by θ. The full data likelihood is:

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n}\prod_{t=a_i+1}^{\tau}\left[f(z_{it}|z_{it-1})\cdot\left(\phi_{it-1}^{s_{it}}(1-\phi_{it-1})^{1-s_{it}}\right)^{s_{it-1}}\left(p_{it}^{\omega_{it}}(1-p_{it})^{1-\omega_{it}}\right)^{s_{it}}\right]. \tag{2.3}$$

This expression follows directly from the EM implementation of Van Deusen (2002) for the use of discrete covariates with one change. Van Deusen (2002) includes an extra exponent on the covariate terms, $f(z_{it}|z_{it-1})^{s_{it}}$, so that only values of z_{it} before each animals death affect the likelihood. The likelihood in equation 2.3 includes values of z_{it} for $a_i < t \leq \tau$ for all individuals. While this may seem peculiar, I regard these extra covariates as values that might have been realized if the animal had continued to live. In truth, this simply represents a larger completion of the data set and has no effect other than to simplify some of the calculations here in the EM and later in the Bayesian approach.

Along with removing the complex multi-dimensional integrals, completing the data allows me to break the likelihood into three components each containing parameters from only one part of the model. Taking the logarithm of equation 2.3, the CDLL can be written as the sum of separate log-likelihood contributions for the models of the covariate drift, survival, and capture:

$$l_z = \sum_{i=1}^{n} \sum_{t=a_i+1}^{\tau} \log f(z_{it}|z_{it-1}) \tag{2.4}$$

$$l_s = \sum_{i=1}^{n} \sum_{t=a_i+1}^{\tau} s_{it}s_{it-1} \log \phi_{it} + (1 - s_{it})s_{it-1} \log(1 - \phi_{it}) \tag{2.5}$$

$$l_c = \sum_{i=1}^{n} \sum_{t=a_i+1}^{\tau} s_{it}\omega_{it} \log p_{it} + s_{it}(1 - \omega_{it}) \log(1 - p_{it}). \tag{2.6}$$

By splitting the likelihood like this the E and M steps can be divided into three parts each operating on only one of the model components.

Unfortunately, while the expression of the CDLL as the sum of these three contributions is much simpler than the observed data likelihood, computation of the E-step in this problem turns out to be very cumbersome. Consider first computing the expected value of the covariate contribution, l_z. By linearity the expected value can be simplified into a sum of univariate expectations:

$$E(l_z|\mathbf{Y}, \theta) = \sum_{i=1}^{n} \sum_{t=a_i+1}^{\tau} \log E(f(z_{it}|z_{it-1})|\mathbf{Y}, \theta).$$

However, each of these univariate quantities is in fact quite complicated to compute.

Depending on the exact pattern of missing data, $E(l_z|\mathbf{Y},\theta)$ can have many different forms based on properties of the MVN distribution discussed in section 2.2. Particularly difficult to handle are the expressions arising from long strings of missing values. On top of this, computing the expected values of the other contributions requires finding the expectation of s_{it} on occasions when z_{it} was not observed. Exactly as before, this requires high dimensional integrals to account for all possible values of z_{it} for $t > b_i$ for each individual. These integrals also involve combinations of logistic functions, through ϕ_{it} and p_{it}, and the MVN density, through $f(z_{it}|z_{it-1})$, and analytic solutions do not exist. Because of these difficulties the EM cannot be used directly and instead I consider a variant of the EM designed to simplify the E-step using numerical integration.

2.3.1 Numerical Integration via MCMC

From the work in the previous section it may appear that the EM algorithm has returned to the same problem that motivated its use, how to work with expressions involving complicated integration. In truth, the situation is greatly improved. Recall that in computing the MLE's directly from the observed data it is necessary to solve derivatives of the integral equations. In the E-step of the EM algorithm, similar integrals arise but they need only be evaluated for the specific parameter values given on the previous M-step. This is a much simpler operation than differentiation and while there are no analytic solutions, the integrals can be evaluated numerically using Monte Carlo integration. In particular, I will make use of Markov Chain Monte Carlo (MCMC) techniques through the Metropolis-Hastings (MH) algorithm and its derivatives.

The driving idea behind Monte Carlo integration is quite simple. Suppose that we are interested in computing the expected value of some function, $g(X)$, where X is a random variable or vector with density $p(\cdot)$. Assuming that it is possible to simulate observations $x_1, \ldots, x_M$ from $p(\cdot)$, the intuitive estimate is:

$$\widehat{E(g(X))} = \frac{1}{M}\sum_{m=1}^{M} g(x_m)$$

This estimator is unbiased and also consistent by the weak law of large numbers. In the words of Gilks et al. (1996): "Monte Carlo integration draws samples from the required distribution, and then forms sample averages to approximate expectations". Of course, how the sample is generated from the arbitrary distribution $p(\cdot)$ is not so obvious.

One set of methods that has been developed to generate such samples is MCMC. Simply put, the goal in MCMC is to construct a Markov chain, X_t, whose stationary distribution matches the desired target, $p(\cdot)$. To do this one must find an appropriate transition kernel, $P(A|X_t = x_t)$, defining the probability that $X_{t+1} \in A$ given that $X_t = x_t$. One method of finding such a transition kernel leads to the MH algorithm. Following the development of Chib and Greenberg (1995), suppose that the kernel is dissected into two components such that:

$$P(A|X_t = x) = \int_A q(y|x)dy + r(x)I[x \in A]$$

Here $q(y|x)$ is the conditional density of the kernel at $y \neq x$, with the stipulation that $q(x|x) = 0$, and $r(x) = 1 - \int q(y|x)dy$ is the probability that the chain remains at x. The Markov chain generated from this kernel will have invariant distribution $p(\cdot)$ if it satisfies the reversibility (Chib and Greenberg, 1995) or detailed-balance (Gilks et al., 1996):

$$p(x)q(y|x) = p(y)q(x|y) \tag{2.7}$$

If one is lucky enough to find such a function, $q(\cdot|\cdot)$, then the problem is solved immediately, but this is unlikely for complex models. Instead, the MH algorithm begins with an arbitrary conditional density and modifies it to satisfy equation 2.7 by introducing a new probability function that regulates how often the possible transitions occur. Suppose that for some pair x and y reversibility is not satisfied because the left-hand side of equation 2.8 is too large:

$$p(x)q(y|x) > p(y)q(x|y) \tag{2.8}$$

Heuristically, this is interpreted as indicating that moves from x to y occur too often relative to transitions in the opposite direction (Chib and Greenberg, 1995). To

balance this, the left side is multiplied by $0 \leq \alpha(y|x) \leq 1$ and the right by $0 \leq \alpha(x|y) \leq 1$ to generate the equation:

$$p(y)q(y|x)\alpha(y|x) = p(x)q(x|y)\alpha(x|y).$$

This equation has an infinite number of solutions and is constrained by settings $\alpha(x|y) = 1$ which yields the unique solution:

$$\alpha(y|x) = \frac{p(x)q(y|x)}{p(y)q(x|y)} \tag{2.9}$$

Starting from the opposite inequality in expression 2.8 one reaches the same result with x and y interchanged. In the terminology of the MH algorithm, $q(\cdot|\cdot)$ is called the proposal (Gilks et al., 1996) or candidate (Chib and Greenberg, 1995) distribution, $\alpha(\cdot|\cdot)$ is called the acceptance probability, and equation 2.9 is called the Hastings ratio. The Markov chain generated from the kernel combining the chosen proposal distribution and the Hastings ratio is guaranteed to converge to its invariant distribution, the target, under further weak conditions (Chen et al., 2000).

Execution of the MH algorithm to generate a sample from the target distribution can be summarized in three steps. Starting from the k^{th} iteration the next element in the chain is generated by:

1. Selecting a proposal y from $q(y|x_k)$.

2. Computing Hastings ratio for y and x_k.

3. Setting $x_{k+1} = y$ with probability $\alpha(y|x_k)$, and $x_{k+1} = x_k$ otherwise.

These steps are then repeated until the chain converges and effects due to the choice of the initial values, x_0, are filtered out. At this point, observations from the Markov chain will be distributed according to $p(\cdot)$ (Carlin and Louis, 2000).

A common method of simplifying the MH algorithm is to work separately on different blocks of variables. If X is a vector quantity then it may be possible to divide it into a set of components $\{X_1, \ldots, X_B\}$, not necessarily all scalars, and update the components one at a time. Making use of the notation of Gilks et al. (1996) I will

denote the set of all components excluding X_b as X_{-b}. Component X_b is updated by drawing a sample from a proposal distribution conditional on the current values of all other components, $q(y|X_b = x_b, X_{-b} = x_{-b})$, and accepting or rejecting based on the component-wise Hastings ratio. From the previous argument it follows that for each component the Markov chain converges to the stationary distribution $p(x_k|x_{-k})$, but it can also be shown that the full vector X updated in this manner also converges to the target distribution (Chib and Greenberg, 1995). This modification is called the block-at-a-time algorithm by Chib and Greenberg (1995) or the single component MH by Gilks et al. (1996).

Choosing the proposal distribution appropriately can also greatly simplify the algorithm. A common choice is to use a proposal that does not depend on the current value of the component being sampled so that $q(y|x_b, x_{-b}) = q(y|x_{-b})$. This type of chain is referred to as an independence sampler (Gilks et al., 1996; Chib and Greenberg, 1995). Gilks et al. (1996) notes that a convenient independence sampler often used is the normal distribution chosen to match $p(\cdot)$ in both spread and location. Another popular method of generating proposals is to sample directly from the full conditionals of each component (Gilks, 1996). For component X_b the full conditional is defined as $p(x_b|x_{-b})$ and can be found in most cases simply by isolating the terms of $p(\cdot)$ involving x_k. When using the full conditional as the proposal, all terms in equation 2.9 cancel leaving $\alpha(y|x_b, x_{-b}) = 1$ for all y so that no proposal is ever rejected. In the specific case where it is possible to sample from the full conditionals for all components of X the algorithm is known as Gibbs sampling and has been explored extensively. Otherwise, it is common practise to sample from the full conditionals when possible and use other proposal distributions for the more complex components. This is commonly termed the MH-within Gibbs algorithm, though reversing this name gives a much more accurate description of its true nature (Chib and Greenberg, 1995; Carlin and Louis, 2000).

Choice of the proposal distribution should also be motivated by how well the resulting chain samples from the target distribution. Poor choices can result in chains that become stuck at some values in the sample space while entirely missing others. Particularly important in determining how well the chain mixes is the variance of the

proposal distribution relative to the target (Chen et al., 2000). If the variance of the chosen distribution is too large then many values will be proposed that have small density under $p(\cdot)$ and the chain will move slowly. Similarly, if the relative variance is too small then many iterations will be needed reach the values where the density is low, usually in the tails of the distribution. Both of these situations lead to inefficient chains that must be run for longer periods in order to compute accurate estimates.

Along with choosing an appropriate proposal several other considerations must be accounted for in any implementation of MH. Most critical are determining the number of samples required to cover the entire distribution and the burn-in period needed to reach convergence (Gilks et al., 1996). Sampling can also be improved by thinning the chain and keeping only observations at evenly spaced steps. This limits the dependence between observations caused by the correlation structure inherent in the construction of the Markov chain. These issues are discussed in depth by Raftery and Lewis (1996) along with introduction of the program `gibbsit` which uses a sample chain to suggest appropriate values for the chain length, burn-in, and sampling step size. Further concerns arise in choosing appropriate starting values and deciding on the optimal order for updating the components in the block-at-a-time algorithm (Gilks et al., 1996). In both the proceeding simulations and data analysis these issues are dealt with very simplistically. Convergence of the chain and burn-in time are both judged visually using trace plots of the iterations and starting values are chosen arbitrarily with little concern for efficiency. These matters should be addressed with more concrete solutions in serious applications.

2.3.2 Monte Carlo Integration and the EM-algorithm

Using the MH algorithm to approximate the expected values in the E-step results in a Monte Carlo EM (MCEM) algorithm. This approach originated from work by Wei and Tanner in the early 1990's and has specific application in problems where the E-step is too complex to compute (Deltour et al., 1999). In a problem similar to this, Ibrahim et al. (1999) use the MCEM to introduce continuous covariates into logistic regression models when some covariate values are missing. The capture-recapture

scenario is distinguished by the fact that for individuals with missing covariates, the response variable is also missing.

In application to the CJS drift model, the MCEM is very simple to implement. Recall that parameters are updated on the M-step and assumed to be known constants on the E-step. To estimate the expected CDLL, I use the MH algorithm to sample complete data sets consistent with both the observed data and the current parameter values and compute the average likelihood. Before proceeding the CDLL is simplified by introducing the statistic $d_i = \max\{t = a_i, \ldots, \tau : \quad s_{it} = 1\}$ for each individual. This statistic represents the last time that individual i was alive and forms a sufficient statistic for the vector $\mathbf{s_i}$ conditional on the first capture. The new complete data likelihood is:

$$\text{Ł}(\theta; \mathbf{x}) = \prod_{i=1}^{n} \left[\prod_{t=a_i+1}^{\tau} f(z_{it}|z_{i,t+1}) \cdot \prod_{t=a_i}^{d_i-1} \phi_{it}(1-\phi_{id_i}) \cdot \prod_{t=a_i+1}^{d} \left(p_{it}^{\omega_{it}}(1-p_{it}^{\omega_{it}})\right) \right] \tag{2.10}$$

and the new CDLL contributions for the models of the covariate, capture, and survival are:

$$\begin{aligned} l_z &= \sum_{i=1}^{n} \sum_{t=a_i+1}^{\tau} \log f(z_{it}|z_{it-1}) \\ l_s &= \sum_{i=1}^{n} \sum_{t=a_i+1}^{d_i} \log \phi_{it} + \log(1-\phi_{id_i}) \\ l_s &= \sum_{i=1}^{n} \sum_{t=a_i+1}^{d_i} (\omega_{it} \log p_{it} + (1-\omega_{it}) \log(1-p_{it})) \end{aligned}$$

From these expressions the likelihood splits so that sampling can be performed independently for each individual. I also use the block-at-a-time algorithm separating the data for each animal into the three basic components: $\mathbf{z_i}$, ω_i, and $\mathbf{s_i}$. Because ω_i is completely known this vector does not need to be sampled. Moreover, on each iteration of the MH algorithm sampling the entire vector $\mathbf{s_i}$ is replaced by sampling the single value d_i.

In equation 2.10 some element of $\mathbf{z_i}$ occurs in every term of $L(\theta; \mathbf{x})$ and it follows that the full conditional of $\mathbf{z_i}$ is proportional to the entire complete data likelihood.

This expression does not correspond with a known distribution and it cannot be sampled directly, so I use an independence sampler instead. Following the recommendations of Gilks et al. (1996), values for the missing covariates are proposed from the multivariate normal distribution conditional on the observed covariate values with mean and variance determined by the current parameter estimates. This approach also seems highly sensible because the proposal matches exactly with the assumed model of the covariate values. Given that the drift parameters found in the previous M-step are $\mu_1, \ldots, \mu_{\tau-1}$ and σ^2, the proposal distribution for $\mathbf{z}$ is multivariate normal with mean vector and variance matrix exactly as described in section 2.2. The corresponding Hastings ratio for a proposed value of the covariates, $\mathbf{z}'$, given the current value, $\mathbf{z}$, is:

$$\alpha(\mathbf{z}'|\mathbf{z}) = \frac{[\prod_{t=a}^{d-1} \phi(z'_t) \cdot (1-\phi(z'_d))] \cdot [\prod_{t=a+1}^{d} p(z'_t)^{\omega_t}(1-p(z'_t))^{(1-\omega_t)}]}{[\prod_{t=a}^{d-1} \phi(z_t) \cdot (1-\phi(z_d))] \cdot [\prod_{t=a+1}^{d} p(z_t)^{\omega_t}(1-p(z_t))^{(1-\omega_t)}]}$$

Perhaps surprisingly, this expression has a very simple interpretation. The new covariate values generated from the model of covariate drift will be accepted with higher probability if they fit more closely with both the current estimates of d and the parameter values as compared with the old values in $\mathbf{z}$. If the new values result in a capture or survival event with very low probability then $\mathbf{z}'$ will be rejected.

Sampling the values of d for each individual is simpler in that d is a discrete random variable. This allows me to sample from the full conditional simply by computing the probability that d takes each value from 1 to τ. For an individual whose current complete covariate vector is $\mathbf{z}$, the full conditional probability $P(d = t|\omega, \mathbf{z})$ for each t is defined as:

$$P(d = t|\mathbf{z}, \omega, \theta) = \begin{cases} 0 & t < b \\ \frac{1-\phi_b}{\chi_b} & t = b \\ \frac{\prod_{r=b+1}^{t}(\phi_{t-1}(1-p_t))\cdot(1-\phi_t)^{I[t<\tau]}}{\chi_b} & b < t \leq \tau \end{cases}$$

The value χ_b in this expression is the probability of not being observed after time b and is simply the sum of the numerators for $t \geq b$. To sample d I generate a value from a $U(0,1)$ random variable and match it with the correct value of t. Because

it is possible to sample from the full conditional of d, this is considered as a Gibbs sampling step within MH.

The biggest drawback in using the MCEM is that the asymptotic properties of the EM are not maintained. Estimates produced will no longer converge point-wise to the MLE's due to the stochastic error introduced in the sampling process. Instead, the algorithm may converge in probability (McLachlan and Krishnan, 1997) or produce a chain whose stationary distribution covers the MLE's (Diebolt and Ip, 1996). For this reason, estimates are usually computed by averaging the results from many iterations after the algorithm appears to have converged in distribution. Once again, the number of E and M-steps required for convergence and the number of iterations needed to provide accurate estimates are dealt with simplistically in the examples. Following the suggestions of McLachlan and Krishnan (1997) trace plots are used to monitor the convergence until the process has stabilized within a certain range of values. More complex methods of monitoring convergence could be incorporated in future applications.

A final difficulty encountered in using the MCEM is calculation of the variance of the parameter estimates. Some techniques have been developed that estimate the variances from a single pass through the algorithm but these are much more involved and instead bootstrapping is often used to estimate the standard errors (McLachlan and Krishnan, 1997). In a non-parametric bootstrap approach, many bootstrap samples are generated by resampling the original dataset, and parameter values are computed for each by repeating the MCEM. Under the assumption that the original data set represents a true random sample from the population, the empirical distribution of the bootstrap estimates can be used to approximate the true distribution. This is the method I use in the following examples.

2.4 A Bayesian Approach

The approach in a Bayesian analysis is fundamentally different to the classical view of statistics. In the classical approach, parameters are viewed as fixed constants unknown to the experimenter; only the data are considered to be random and modelled using

probabilities. Bayesian statistics works on the alternative premise that parameters are random variables no different from data. In the Bayesian paradigm, information from the experiment is used to construct a distribution, called the posterior, that describes how possible values of the parameter rank in plausibility given the observed data. Inference about the parameters involves computing different statistics to determine which parameters or sets of parameters have the highest probability given the observed data. This perspective provides a completely different approach to estimating the parameters in the CJS drift model.

Regardless of the exact problem, Bayesian analysis follows a strict recipe combining two ingredients. The first is the prior distribution which describes the experimenter's knowledge about the parameter values before the experiment begins. This distribution can be chosen in many different ways, as will be discussed, but in general it is up to the experimenter to choose a distribution that ranks the values in the parameter space according to her subjective knowledge about the system. Extending the notation of the previous sections, θ will denote the set of all parameters and $\pi(\theta)$ the prior distribution. Data gathered during the experiment, denoted by $\mathbf{x}$, is incorporated into the analysis through the second ingredient, the likelihood function, $L(\theta; \mathbf{x})$. Conditioning on the parameter values, these two functions can be multiplied to define a joint distribution for both the data and the parameter values:

$$p(\mathbf{x}, \theta) == p(\mathbf{x}|\theta)\pi(theta)L(\theta; \mathbf{x})\pi(\theta) \tag{2.11}$$

To determine the posterior distribution of the parameters conditional on the data, $\pi(\theta|\mathbf{x})$, Bayes' theorem is used to reverse the conditioning in expression 2.11 (Carlin and Louis, 2000):

$$p(\theta|\mathbf{x}) = \frac{p(\theta, \mathbf{x})}{\int p(\theta, \mathbf{x}) d\theta} = \frac{\pi(\theta)L(\theta; \mathbf{x})}{\int \pi(\theta)L(\theta; \mathbf{x}) d\theta}$$

In most cases, the normalizing constant in the denominator can be dropped and the posterior is uniquely defined by the proportionality statement:

$$p(\theta|\mathbf{x}) \propto \pi(\theta)L(\theta; \mathbf{x})$$

This statement also summarizes very nicely the fundamental idea behind all Bayesian

analysis. In essence, previous knowledge from the prior is combined with new information gained during the experiment using Bayes' theorem to condition on the observed data. As more experiments are performed, the process can be repeated starting with the updated prior, leading to a succession of more informative posterior distributions.

The single step that is often the most difficult in a Bayesian analysis is construction of a reasonable prior distribution before conducting the experiment. This process is also strongly criticized by frequentists who challenge that the prior distribution represents subjective views about the parameters and that different priors may lead to very different conclusions. In most situations the prior distribution is actually built from information gained during previous studies or pilot projects, and reflects expert knowledge about the underlying physical or biological system. Moreover, in cases where no previous information is available it is common to use a non-informative prior distribution to limit the effects on the final inference (Carlin and Louis, 2000). This can be done by choosing a distribution which varies widely over the parameter space without considerable mass or density at any specific point and in the extreme case produces a flat prior with equal probability on each point in the parameter space. For continuous parameter spaces such distributions are called improper priors because it is impossible to construct a density equal at all points which integrates to 1. However, setting $\pi(\theta) = c$ for any $c \in R$ in equation 2.4 results in a cancellation such that the posterior is simply proportional to $L(\theta; x)$.

Another method used for constructing priors is to choose distributions that simplify the calculations needed to make inference. One way of doing this is to choose a prior distribution that results in a well known posterior distribution. Particularly appealing are the classes of conjugate priors for which the prior and posterior both belong to the same distributional family (Carlin and Louis, 2000). While choosing the prior for computational simplicity rather than to reflect the experimenter's true beliefs is frowned upon by strict Bayesian statisticians, using conjugate priors is a very common practise.

In the simulations and analysis presented in chapter 3, I assume that no previous information is available about any of the parameters and use non-informative priors. The prior distribution is also broken into three independent components for the models

of covariate drift, capture, and survival. In total, the drift model contains τ different parameters, $\mu_1, \ldots, \mu_{\tau-1}$ and σ^2 which I also assume to be independent. For the drift mean, μ_t, I use a normal prior with mean 0 and large prior variance, σ_μ^2 so that the density is spread as widely and evenly as possible:

$$\pi(\mu_i) \propto \exp\left\{-\frac{\mu_t^2}{2\sigma_\mu^2}\right\}$$

This distribution is also the conjugate prior for the drift means. Similarly, the conjugate family for σ^2 is the inverse gamma distribution with parameters α_σ and β_σ:

$$\pi(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\alpha_\sigma+1} \exp\left\{-\frac{1}{\beta_\sigma\sigma^2}\right\}$$

The hyper-parameters α_σ and β_σ are chosen so that the mean of this distribution is 1 and variance is again very large. The remaining parameters in the model are the coefficients in the logistic regression equations for capture and survival and computations are greatly simplified if improper flat priors are used such that $\pi(\alpha_0, \alpha_1)$ and $\pi(\beta_0, \beta_1)$ are uniformly equal to 1. Following the assumption of independence between the three sets of parameters the complete prior distribution is specified by:

$$\begin{aligned} \pi(\theta) &= \prod_{t=1}^{\tau-1} \pi(\mu_t) \cdot \pi(\sigma^2) \cdot \pi(\beta_s)\pi(\beta_c) \\ &\propto \exp\left\{-\frac{\sum_{t=1}^{\tau-1}\mu_t^2}{2\sigma_\mu^2} - \frac{1}{\beta_\sigma\sigma^2}\right\} \cdot \left(\frac{1}{\sigma^2}\right)^{\alpha_\sigma+1} \end{aligned} \tag{2.12}$$

Of course, any information that is available about these parameters should be incorporated to make priors specific to each application. The likelihood does not change from the frequentist approach and remains exactly as in 2.10.

While the posterior distribution contains all information about the parameters, it is usually too complicated to interpret directly. Instead, summary statistics are used to describe the posterior and indicate which parameter values have the highest probability. Carlin and Louis (2000) suggests a summary involving 3 important components. First, point estimates are given for each parameter to indicate the centre of the posterior distribution. Commonly either the mode or the mean of the posterior

are given as point estimates. I have chosen to use the mean because it minimizes the posterior variance with respect to $\hat{\theta}(x)$, $V = E_{\theta|X}[(\hat{\theta}(X) - \theta)^2]$ over all estimators $\hat{\theta}(X)$ (Carlin and Louis, 2000). Interval estimates are used to judge the precision of these estimates and the variance of the posterior distribution. In Bayesian statistics these are called credible sets, and a $(1 - \alpha)100\%$ credible set is any subset $C \in \Theta$ satisfying:

$$P(\theta \in C) \geq 1 - \alpha$$

Ideally, C should be chosen so that it contains the values most probable under $\pi(\theta|\mathbf{x})$, i.e. those with the highest posterior density (HPD), but this can be difficult to compute and equal tail intervals are used instead (Carlin and Louis, 2000). Finally, Carlin and Louis (2000) suggest providing plots of the posterior density to convey the major features of the distribution.

In reporting the conclusions of the simulations and analysis in sections 3.1 and 3.2 I have included point estimates, interval estimates, and plots where appropriate. Plots of the distributions are somewhat revealing for the CJS drift model, but the high dimensionality and correlation between some of the parameters make them difficult to interpret. As in the majority of Bayesian analyses, evaluation of the point and interval estimates is completely impossible (Gilks et al., 1996). Exactly as in the EM algorithm, problems arise in computing expected values that involve high dimensional integrals and can only be computed numerically. For this, I turn again to MCMC methods and the MH algorithm.

2.4.1 Monte Carlo Estimation for the Bayesian Approach

To sample from the Bayesian posterior distribution I again use the block-at-a-time MH method. Because parameters behave as random variables, several more blocks must be updated on each iteration. Along with the data components for each individual described in 2.3.2, the parameters are broken into components for the covariate drift means, drift variance, and the regression coefficients of the capture and survival models. In total, each iteration of the MH algorithm requires 6 separate updates.

Solving for the full conditionals of the missing data variables from the Bayesian posterior distribution shows that they are exactly as described in section 2.3.2, and hence, the methods of sampling $\mathbf{z}$ and d for each individual are unchanged. It remains to define steps to simulate the model parameters.

After generating a complete data-set for each individual the drift means and variance are the first to be updated. For the vector of drift means, μ, the full conditional isolated from the posterior distribution is given by:

$$\begin{aligned}\pi(\mu|X,\theta_{-\mu}) &\propto \prod_{i=1}^{n}\prod_{t=a_i+1}^{\tau} f(z_{it}|z_{it-1})\cdot\pi(\mu) \\ &= \prod_{t=1}^{\tau-1}\left[\prod_{i=1}^{n} f(z_{it+1}|z_{it})^{I[a_i<t]}\cdot\pi(\mu_t)\right]\end{aligned}$$

The assumption of independence made in defining the prior has carried over to the posterior so that each μ_t can be generated independently. Moreover:

$$\mu_t|X,\theta_{-\mu} \sim N\left(\frac{\sum(z_{it+1}-z_{it})+\frac{\sigma^2}{\sigma_\mu^2}\mu_{t0}}{n+\frac{\sigma^2}{\sigma_\mu^2}} \quad , \quad \frac{\sigma^2}{n+\frac{\sigma^2}{\sigma_\mu^2}}\right)$$

a univariate normal distribution that is easily simulated using any statistical package. For σ^2 the full conditional distribution is given by the expression:

$$\pi(\sigma^2|X,\theta_{-\sigma^2}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\sum_{i=1}^{n}(\tau-a_i)}{2}+\alpha_0+1}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{\beta_0}+\frac{\sum_{i=1}^{n}\sum_{t=a_i}^{\tau-1}(z_{it+1}-z_{it}-\mu_t)^2}{2}\right)\right\}$$

This shows that conditional on the other variables, the variance follows an inverse gamma distribution:

$$\sigma^2|X,\theta_{-\sigma^2} \sim IG\left(\frac{\sum_{i=1}^{n}(\tau-a_i)}{2}+\alpha_\sigma \quad , \quad \left(\frac{1}{\beta_\sigma}+\frac{\sum_{i=1}^{n}\sum_{t=a_i}^{\tau+1}(z_{it+1}-z_{it}-\mu_t)^2}{2}\right)^{-1}\right)$$

Again, samples from this distribution are easily generated either directly or by inverting a simulated value from the corresponding gamma distribution. Both of these cases illustrate the advantage of using conjugate priors. Carefully choosing the priors has led to convenient full conditionals.

Sampling the coefficients of the regression models is not so simple. The full conditional distributions that result are quite complicated and cannot be used directly. As with the missing covariates I turn to an independence sampler approach based on a MVN proposal distribution. For the model of the capture process, the location and spread of the proposal distribution are chosen by conducting a logistic regression using the current complete capture indicators for each individual as the response and the completed covariate data as the dependent variable. The resulting vector of coefficients is used as the multivariate normal mean and the spread is determined by the associated inverse information matrix. For an experiment involving only a single covariate this defines a bivariate normal distribution from which a new proposal for the coefficients is selected and accepted or rejected based on the Hastings ratio. Letting β_c represent the current coefficients, $\hat{\beta}$ the coefficients resulting directly from the logistic regression with observed information matrix $I(\hat{\beta})$, and β_c' the new proposal, the Hastings ratio is for the new proposal is:

$$\alpha(\beta_c'|\beta_c) = \frac{\prod_{i=1}^{n}\prod_{t=a_{i+1}}^{d_i}\left[(p_{it}')^{\omega_{it}}(1-(p_{\omega_i t}'))^{1-\omega_{it}}\right]\cdot\exp\left\{-\frac{1}{2}(\beta_c'-\hat{\beta}_c)^T I(\hat{\beta}_c)(\beta_c'-\hat{\beta}_c)\right\}}{\prod_{i=1}^{n}\prod_{t=a_{i+1}}^{d_i}\left[p_{it}^{\omega_{it}}(1-p_{\omega_{it}})^{1-\omega_{it}}\right]\cdot\exp\left\{-\frac{1}{2}(\beta_c-\hat{\beta}_c)^T I(\hat{\beta}_c)(\beta_c-\hat{\beta}_c)\right\}}$$

Here $p_{it}' = p(z_{it};\beta_c')$ and $p_{it} = p(z_{it};\beta_c)$. Once again, the Hastings ratio has a very intuitive interpretation. Both the numerator and denominator can be broken into two terms; the first of these is the likelihood for the capture events, and the second measures the distance between the proposed or current parameters and the mean of the normal distribution. If the proposed coefficients do not fit well with the data such that the likelihood is small or have low density under the proposal distribution, then they will be rejected with high probability. Similarly, if the magnitude of $\beta - \hat{\beta}$ is large then the proposal will be accepted almost always. Exactly the same procedure is used to sample the coefficients of the survival model. In sampling the regression coefficients for both capture and survival it is important that the vector of coefficients is simulated in a single step and not broken into smaller components. Within the two logistic models, the regression parameters are highly correlated and sampling one coefficient conditional on the other will lead to a chain that moves slowly taking many more iterations to sufficiently cover the sample space.

Using this MCMC setup the important Bayesian summary statistics can now be computed. Alternating between the 6 steps, the Markov chain is iterated until convergence. Point estimates are provided by the posterior means of each parameter approximated by averaging over many observations from the tail of the chain. Interval estimates are also computed using the observed quantiles of the tail to approximate equal-tailed credible sets.

Chapter 3

Applications

3.1 Analysis of Simulated Data

In the course of this project I analyzed many different data sets in order to compare and validate the estimation techniques. I have included the results from two specific examples. The first demonstrates the use of the CJS drift model under ideal conditions, and both methods produce very good estimates of the underlying survival and capture processes for this data set. In contrast, the second data set has a change in the capture coefficients that results in a much larger proportion of missing data and poses a much bigger challenge. Full data MLE's are used as a benchmark to judge the performance of the two methods.

Both simulations described used exactly the same model of the covariate drift. A total of n=200 individuals were included in each data set with $\tau = 5$ different capture occasions. For each individual the value of the covariate at time $t = 1$ was randomly selected from a normal distribution with mean 100 and variance 5 and the remaining values were generated from the MVN drift model with mean vector $\mu = (10, -1, 1, -10)$. Keeping with the analogy of body mass, I chose these values to suggest a population with a very good initial year in which the average animal gains 10% of its body mass, followed by a weakly poor year, a weakly good year, and a very poor year. The constant drift variance was $\sigma^2 = 5$. This value allowed some animals to go against the trend in the weak years, but in the very good year it was

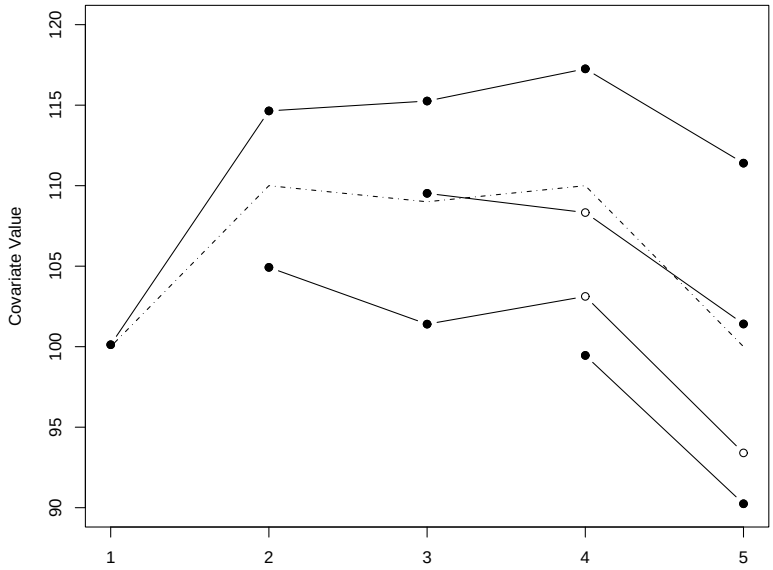

Figure 3.1: Sample covariate values for 4 individuals from the first simulation, each with a different first capture occasion. Solid points represent occasions when the animal was captured and the covariate was recorded, open circles represent occasions when the animal was not captured. The broken line shows the mean trend in the covariate value.

unlikely for an individual to lose weight and in the very poor year it was unlikely to gain weight. In an attempt to maintain a fairly constant sample size across all capture periods I arranged the animals so that 50 were first captured on each of $t = 1, 2, 3, 4$. This avoided problems that arose in some simulations when very few animals were captured on the final occasions because no new animals had been introduced into the population since the first capture. Sample histories for 4 individuals from the first simulation are shown in Figure 3.1.

The exact formulation of the MCEM and Bayesian MH estimation schemes was also common to both simulations. The Bayesian analysis ran for a total of 5000 iterations and the parameter estimates from the final 500 were kept as a sample from

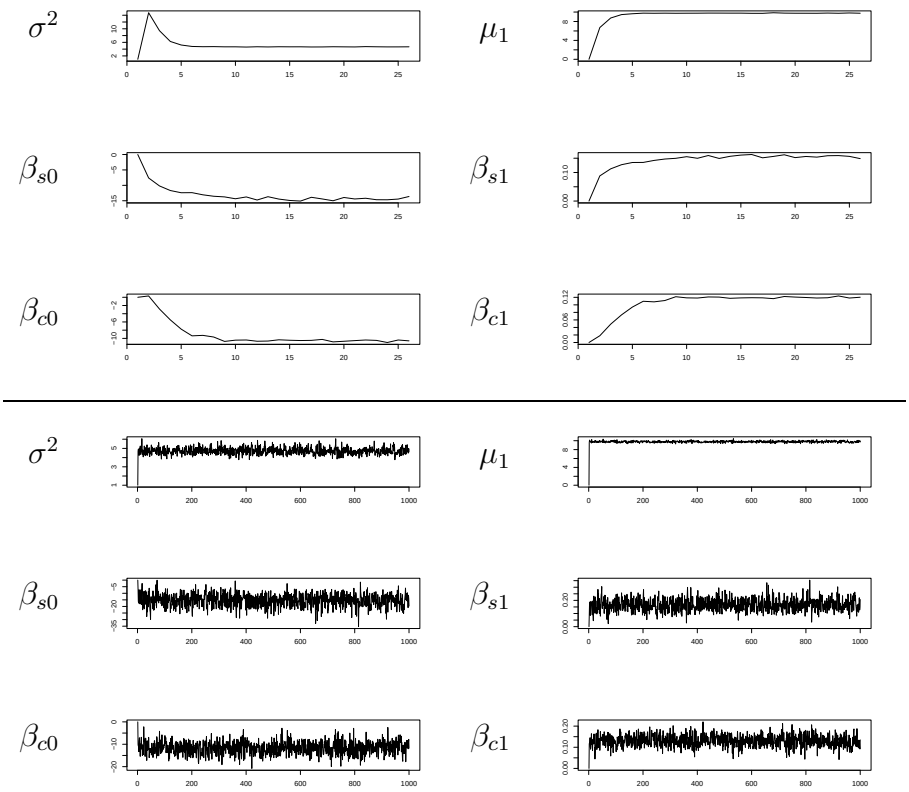

Figure 3.2: Trace plots of the estimates of selected model parameters for each iteration of the MCEM (top) and the Bayesian MH algorithm (bottom) from the first simulation.

the posterior distribution. The MCEM ran for 25 EM iterations and each E-step consisted of 50 MCMC iterates. Parameters produced on the final 5 EM iterations were averaged to compute the parameter estimates. To check the convergence of the algorithm, I produced trace plots of the parameter estimates for each iteration of the MCEM and Bayesian methods, and these are included for selected parameters for the first simulation (Figure 3.2). From these plots it is evident that the parameter estimates converged quickly for both methods reaching a stable distribution in the given number of iterations. Finally, for the MCEM 50 bootstrap simulations were used to compute the variance estimates.

	Full Data MLE	Bayesian MH	MCEM
$\mu_1 = 10.00$	9.86	9.77	9.76
	(9.13,10.5)	(9.38,10.17)	(9.19,10.38)
$\mu_2 = -1.00$	-0.99	-1.14	-0.96
	(-1.46,-0.50)	(-1.20,-0.77)	(-1.49,-0.45)
$\mu_3 = 1.00$	0.94	0.88	0.892
	(0.56,1.32)	(0.71,1.05)	(0.40,1.32)
$\mu_4 = -10.00$	-9.70	-9.78	-9.71
	(-10.05,-9.36)	(-9.98,-9.56)	(-10.16,-9.49)
$\sigma^2 = 5.00$	4.63	4.72	4.68
	(3.84,5.70)	(4.05,5.50)	(3.93,5.10)
$\beta_{s0} = -10.00$	-15.25	-9.70	-14.28
	(-21.23,-6.31)	(-26.19,-6.77)	(-30.64,-3.90)
$\beta_{s1} = .115$	0.150	0.165	0.155
	(.079,.220)	(.080,0.268)	(.053,.276)
$\beta_{c0} = -10.00$	-8.83	-11.12	-10.59
	(-13.80,-3.85)	(-17.14,-5.89)	(-20.31,-6.00)
$\beta_{c1} = .115$	0.103	0.126	0.112
	(0.055,0.151)	(0.073,0.184)	(0.063,0.202)

Table 3.1: Results of the first simulation experiment. The first column gives the true parameter values followed by the estimates from the different estimation methods. Below each estimate are provided 95% confidence intervals for the full data MLE's and MCEM estimates, and 95% credible intervals for the Bayesian estimates.

To gauge the performance of the different methods I calculated the full data MLE's for both simulated data sets. These values are simply the estimates of the drift parameters and regression coefficients computed using full knowledge of the data including the unobserved covariates and actual survival indicators. These estimates provide a target for the two new methods and gauge how accurate and precise a good estimate could possibly be.

In the first simulation I chose the regression coefficients for the survival and capture curves specifically so that the curves would be easy to recover. Both sets of coefficients were the same, $\beta_c = \beta_s = (-10, 0.115)$ and the resulting curves, shown in Figures 3.3 & 3.4, share two important characteristics. First, the curves show a very

strong trend over the observed range of the covariate that I hoped would be easily detected and differentiated from a constant rate. Further, the values of both curves remain above 0.5 throughout the entire range coming very close to 1 at the upper limit. This meant the majority of animals were likely to survive and to be captured on several occasions, and in fact those at the upper limit would be captured almost surely. In total, 551 captures occurred out of a possible 1000.

Point estimates and interval estimates for the first simulation are reported in Table 3.1. Plots of the estimated curves with approximated 95% confidence/credible envelopes are shown in Figures 3.3 & 3.4. Both methods yielded very accurate estimates and the point estimates lie close to the true values for all parameters both in absolute value and relative to the full data MLE's. Moreover, with the exception of a single case the interval estimates all cover the true parameter values and in general are well balanced on either side. The only interval that fails to do so is the Bayesian 95% credible interval for μ_4. However, 9 credible intervals were generated simultaneously and so with probability possibly as low as $1 - 0.95^9 = 0.37$ at least one true value will fall outside of the associated interval estimate. This failure is likely due to chance, and the interval does in fact come very close to covering the true value of μ_4, missing by only 0.02 in the lower limit.

Precision of the estimates is gauged by the width of the corresponding intervals. In all cases, the intervals are sufficiently narrow to show that the parameters are significantly different from 0 at the 5% level, without correction for multiple comparisons. From this one would conclude that there is significant effect of the covariate on both capture and survival and that the drift is different in each of the 5 years. The intervals produced from the MCEM and Bayesian MH were relatively narrow in comparison to the interval estimates for the full data MLE's indicating little loss in precision due to the missing data. This is also visible in the plots of the final curves where the confidence/credible envelopes easily cover the true curves and are only slightly wider than the confidence envelopes for the full data MLE's. Not surprisingly, the methods performed particularly well at the upper limit of the observed covariate values where most captures occur. Unfortunately, even though it does give some indication of the performance, comparing the intervals and envelopes for the

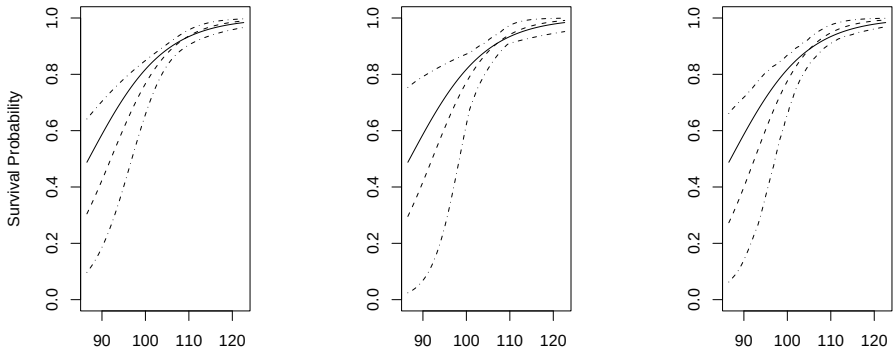

Figure 3.3: Estimated survival curves for the first simulation using the three estimation methods: full data MLE's (left), MCEM (centre), and Bayesian approach (right). The solid line in each plot shows the actual survival curve and the dashed line the estimated curve. The dot-dash lines in the figures give the 95% confidence/credible intervals for the value of $\phi(z)$. The observed range of covariate values was $86.5 \leq z \leq 122.7$.

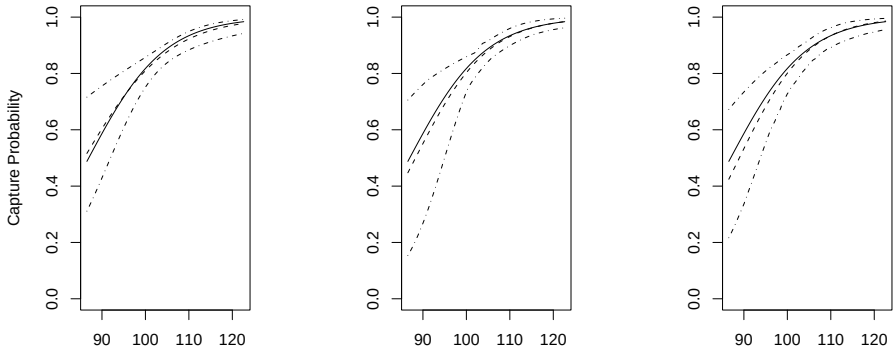

Figure 3.4: Estimated capture curves for the first simulation using the three estimation methods: full data MLE's (left), MCEM (centre), and Bayesian approach (right). The solid line in each plot shows the actual capture curve and the dashed line the estimated curve. The dot-dash lines in the figures give the 95% confidence/credible intervals for the value of $p(z)$ for each z in the range of observed values.

Bayesian method with those of the full data MLE's is not strictly appropriate because of the different interpretation for confidence intervals and credible intervals. A more appropriate comparison would use the Bayesian estimates from the full data in place of the frequentist MLE's.

To complete the steps of Bayesian inference recommended by Carlin and Louis (2000) I have included plots of the posterior marginals. Histograms of the sampled drift parameters for σ^2 and μ_1 are shown (Figure 3.5) with the fitted full conditional curve. In both cases, the full conditional fits the histogram very well providing more evidence that the algorithm has indeed converged. Plots of the marginals for the three other drift means were very similar. Histograms of the sampled values of the coefficients of the survival curve are also shown (Figure 3.6). These plots indicate that both of these parameters had skewed distributions. Also important is the very strong correlation between the two coefficients, visible in the centre plot. This correlation is not surprising and was what necessitated sampling $(\beta_s 1, \beta_s 2)$ as a single component in the MH algorithm. Plots for the components of β_c were very similar and are not included.

In the second simulation I modified only the capture parameters. The new vector of coefficients for capture was $B_c = (0.10, -0.0011)$ and the effect of this change is visible in Figure 3.8. The capture curve is almost flat, reaching a maximum near 0.51 and minimum of 0.49 over the observed range of covariates and is also turned so that the capture probability actually decreases as the covariate increases. The major result of this change is to lower the overall capture rate so that less information was available for estimation; only 390 captures occurred in this simulation. Trace plots showed again that the MCEM and Bayesian MH converged using the given schemes and the resulting parameter estimates are shown in Table 3.2. Not surprisingly the estimators did not perform as well as in the first simulation. While the point estimates are reasonably close to the true parameter values and the interval estimates do cover the true values in all but one case, the confidence/credible intervals for the MCEM and Bayesian MH are very wide. I did not expect that either method would show a significant effect of the covariate on capture because the full data MLE's of $\beta_c 1$ is not significantly different from 0. However, both methods do fail to detect the effect on

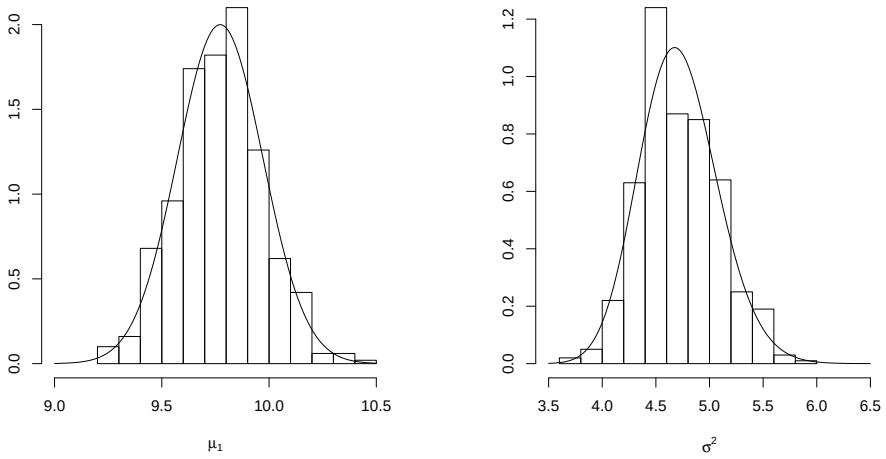

Figure 3.5: Histograms of the sampled values of μ_1 (left) and σ^2 (right) for the Bayesian MH method. The curves indicate the normal full conditional density for μ_1 and inverse Gaussian for σ^2.

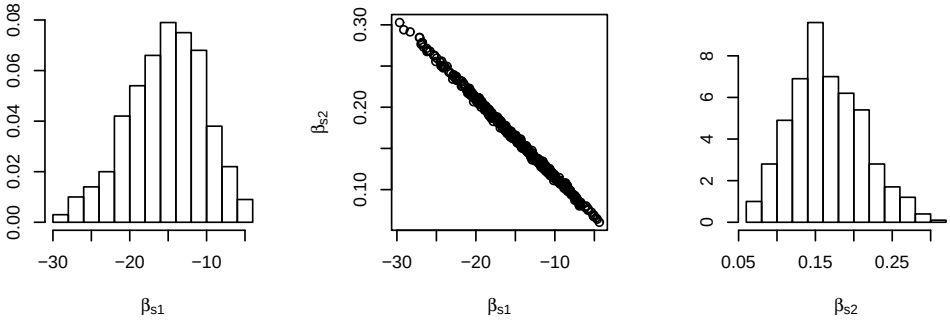

Figure 3.6: Histograms of the sampled values of β_{s0} (left) and β_{s1} (right). The centre panel plot β_{s1} vs. β_{s0} for each of the 500 samples.

	Full Data MLE	Bayesian MH	MCEM
$\mu_1 = 10.00$	9.85	9.75	9.67
	(9.13,10.6)	(9.31,10.16)	(8.64,10.67)
$\mu_2 = -1.00$	-0.98	-0.80	-0.77
	(-1.46,-0.50)	(-1.23,-0.36)	(-1.31,-0.30)
$\mu_3 = 1.00$	0.94	0.91	0.94
	(0.56,1.32)	(0.54,1.30)	(0.34,1.40)
$\mu_4 = -10.00$	-9.70	-9.83	-9.85
	(-10.05,-9.36)	(-10.21,-9.48)	(-10.48,-9.19)
$\sigma^2 = 5.00$	4.63	4.39	4.36
	(3.84,5.70)	(3.64,5.29)	(3.43,4.78)
$\beta_{s0} = -10.00$	-13.77	-8.91	-8.03
	(-21.23,-6.31)	(-22.13,3.92)	(-21.04,0.64)
$\beta_{s1} = .115$	0.150	0.100	0.090
	(.079,.220)	(-0.023,0.224)	(-0.104,0.205)
$\beta_{c0} = 0.100$	1.50	2.80	3.55
	(-1.85,4.85)	(-1.05,7.07)	(-4.55,10.57)
$\beta_{c1} = -0.0011$	-0.015	-0.026	-0.03
	(-0.047,0.016)	(-0.066,0.010)	(-0.104,0.025)

Table 3.2: Results of the second simulation experiment. Below each point estimate are provided 95% confidence intervals for the full data MLE's and MCEM estimates, and 95% credible intervals for the Bayesian estimates.

survival which is recovered by the full data MLE's. This result is also evident in the plots of the estimated survival curves (Figure 3.7) which have very large envelopes that easily fit lines with constant probability.

From the plots and the results of the simulations shown in Tables 3.1 & 3.2 it appears that the MCEM and Bayesian methods perform very similarly. Both produce point estimates that are closer to the true value or have higher precision than the other for some parameters, but neither performs uniformly better. There is, however, one very major advantage to the Bayesian method: the algorithm does not have to be repeated to produce variance estimates. The MCEM on the other hand has to be repeated once for each bootstrap sample. This was very time consuming and limited

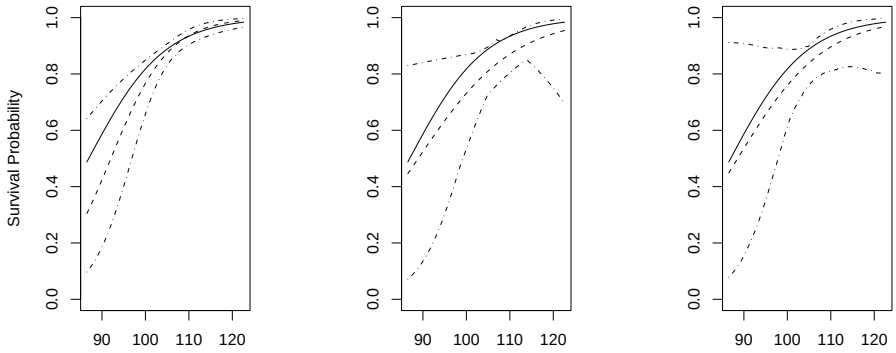

Figure 3.7: Estimated survival curves for the second simulation using the three estimation methods: full data MLE's (left), MCEM (centre), and Bayesian approach (right). The solid line in each plot shows the actual survival curve and the dashed line the estimated curve. The dot-dash lines in the figures give the 95% confidence/credible intervals for the value of $\phi(z)$ for each z in the range of observed values.

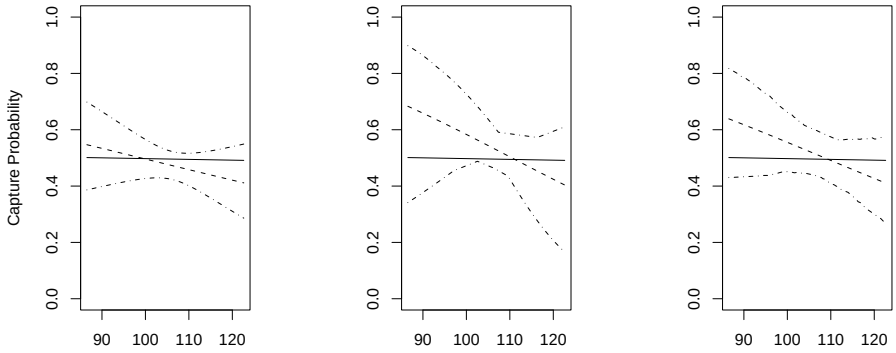

Figure 3.8: Estimated capture curves resulting for the second simulation using the three estimation methods: full data MLE's (left), MCEM (centre), and Bayesian approach (right). The solid line in each plot shows the actual capture curve and the dashed line the estimated curve. The dot-dash lines in the figures give the 95% confidence/credible intervals for the value of $p(z)$ for each z in the range of observed values.

both the number of EM iterations and the number of MCMC iterations on each E-step. In fact, in order to generate accurate estimates of the variances far more than 50 bootstrap samples should have been used, but this was simply not practical. With only 50 samples the precision of the order statistics is very low and the bounds of the 95% confidence intervals should be treated as rough approximations.

3.2 Analysis of Snow Goose (*Chen Caeruslescens*) Weight Data

The sample data set was provided by Dr. Evan Cooch of the Department of Natural Resources, Cornell University. This data contains capture records for 31,240 individual snow geese (*Chen Caeruslescens*) observed over a 19 year period. In total, there are 40,457 captures, an average of 1.30 per individual. Also provided for some recoveries are records of the bird's body mass; length of the culmen, the exposed portion of the beak; and the length of the tarsus, the lower part of the bird's leg.

As an index of the bird's overall health, Cooch recommends the ratio of body mass to culmen length. If the index is based only on body mass then some birds will have a higher index not because they are actually healthier but because they are simply larger individuals. Most notably, the distribution of body mass is markedly bimodal because of a difference in the mean weight of males and females, but this does not mean that the larger males are healthier animals. Culmen length is a measure of the size of the bird's skeleton and remains virtually constant throughout it's adult life. The ratio of body mass to culmen length measures each bird's fatness instead of the absolute size, with the assumption that fatter birds are healthier. For this data set the distribution of the ratio is both very symmetric and unimodal with a mean value near 35 g/mm over all captures.

To run the analysis I found it necessary to remove many records from the data set. Relatively few records contained values for both body mass and culmen length because the researchers often failed to measure these values when the birds were captured. I assumed that these values were missing purely at random and removed the entire

Drift Parameters

μ_4	μ_5	μ_6	μ_7	μ_8	σ^2
0.24	1.25	0.41	-0.58	-0.12	2.40
(-0.29,0.66)	(0.79,1.75)	(0.10,0.82)	(-0.88,-0.30)	(-0.28,0.13)	(2.08,2.72)

Survival and Capture Parameters

β_{s0}	β_{s1}	β_{c0}	β_{c1}
1.13	0.011	-6.20	0.118
(-2.75,7.77)	(-0.16,0.12)	(-7.99,-4.29)	(0.06,0.17)

Table 3.3: Parameter estimates computed using the Bayesian MH algorithm. 95% credible intervals for each parameter are given below the point estimates. The data was chosen from years 4-9 of a much larger study. Hence, μ_4 refers to the mean change between years 4 and 5 in the original study, but between the 1^{st} and 2^{nd} years in the selected data.

ϕ_1	ϕ_2	p_1	p_2
0.763	0.924	0.068	0.166
(0.581,0.883)	(0.678,0.986)	(0.045,0.101)	(0.122,0.221)

Table 3.4: Estimated parameters for the Arnason-Schwarz model. The survival and capture rates for strata k are denoted by ϕ_k and p_k. 95% confidence intervals appear below each estimate.

records without any adjustment. Several birds also had repeated captures in single years with different values for either the body mass or the culmen length. When this occurred I decided to keep only the final record and discard all others. To restrict the size of the data set I also focused on 6 years of data, years 4 through 9. The reduced data set contains a total of 1760 full records for 1511 individuals captured between years 4 and 9. The average number of captures per individual in this subset is 1.17 and no individual has more than 4 complete records.

Based of the results from the simulations I decided to use the Bayesian method first and postpone the MCEM. As in section 3.1, I ran the MH algorithm for a total of 5000 iterations and retained the values from the last 500 as a sample from the

posterior distribution. Trace plots suggest that the algorithm still converged very quickly even with this larger data set (Figure 3.9). Estimates based on the sample are given in Table 3.3 and plots of the resulting capture and survival curves are shown in Figure 3.10.

The most significant result from the analysis is the strong relationship between the health index and capture. Figure 3.10 shows that the capture rate increases with the health of the bird starting near 0.01 for the least healthy and reaching a maximum near 0.25 for the most healthy. Unfortunately, this result must be interpreted in light of the fact that many records were removed from the data set; the capture rate is much more reflective of the rate at which the researchers recorded both measurements than the rate of capture. However, if the measurements in the data set are missing purely at random then this result would still suggest that healthier birds were captured significantly more often. The analysis also reveals that some years showed a significant change in the populations overall health. In particular, years 5 and 6 produced overall gains in health while year 7 produced a significant decline. I have contacted Cooch and hope that he will be able to suggest reasons for these trends. Finally, the estimated survival curve does show increasing rates with increasing health, as expected, but the trend is not significant. In Figure 3.10 the credible envelope is very wide at all points and a constant rate near 0.8 could very plausibly fit all animals. Once again, this might be a result of thinning the data. Unlike the capture rate, if the records truly are missing at random then the interpretation of the survival rate does not change. It is simply as if many fewer capture occurred decreasing the amount of information available to estimate the survival parameters, driving up the standard error of these estimates, and making it more difficult to detect significant trends.

For validation I compared these results with estimates from a 2 strata multi-state Arnason-Schwarz model. Without a priori knowledge of snow goose biology, I divided the health index simply so that the same number of observations occurred in both strata. Estimates of the survival and capture rates for this model were produced using the package MARK developed for capture-recapture analysis by Dr. White of the Department of Fishery and Wildlife Management at Colorado State University. To match the assumptions of the CJS drift model, transition rates were allowed to

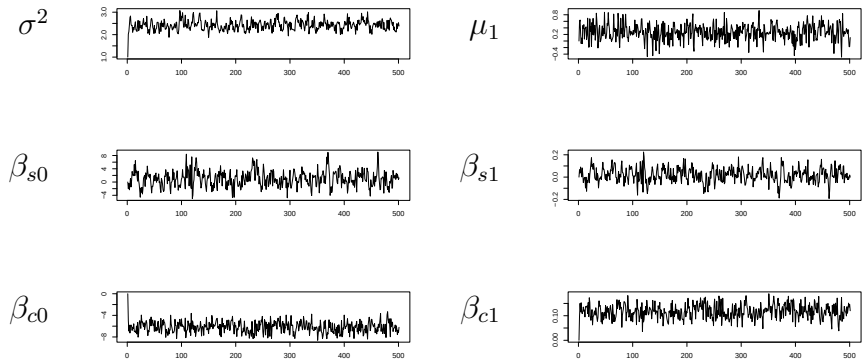

Figure 3.9: Trace plots of the estimated model parameters for each iteration of the Bayesian MH algorithm.

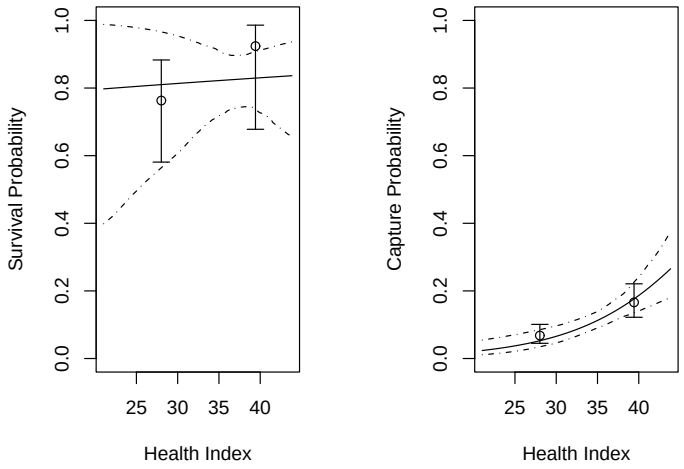

Figure 3.10: Estimated survival (left) and capture (right) curves for the snow goose data. The solid line on each plot shows the estimated curve and the dot-dash lines give 95% credible envelopes for each value of the health index. The two dots in each plot give the estimated rates with 95% confidence intervals for the 2 strata used in the Arnason-Schwarz model and are centred at the middle of each strata.

vary over time while the capture and survival parameters were held constant. The resulting estimates of the survival and capture probabilities, given in Table 3.4 and also visible in Figure 3.10, match very well with the results for the CJS drift model. The capture rates centred at the middle of the two strata lie almost exactly on the estimated continuous curve and confirm the significant increase in rate with increasing health. Although the estimated survival rates do not match as well with the continuous survival curve, the 95% confidence intervals indicate that these estimates are very variable and do easily cover the curve at both locations. Better comparisons and stronger validation might come from a model using more strata, but even a model with 3 strata and constant transition rates could not be estimated by MARK because of problems with parameter identifiability.

As for the simulated data sets I did repeat the analysis using the MCEM method, but this yielded very peculiar results. While most of the parameters converged to stable limits consistent with the Bayesian estimates, the coefficients of the survival curve converged to very different values. Final parameter estimates were well outside of the 95% credible intervals of the Bayesian estimates and the curve produced did not match either the Bayesian estimated logistic curve or the point estimates of the 2 strata Arnason-Schwarz model. Even more peculiar was the fact that exactly the same behaviour resulted when analyzing simulated data sets with similar parameter values. In view of the performance on the simulated data sets in section 3.1, I do not believe that this is the result of an error in the implementation. It is possible that this method of estimation suffers from the same model identifiability problems of the more flexible multi-strata models, but at the moment I am not able to present a clear reason for these results.

Chapter 4

Conclusion

I believe that this project presents a significant contribution to capture-recapture methodology. As noted in the introduction, previous methods for incorporating covariates into the CJS model require that the covariates be discrete, static, or common to the entire population. Continuous time-dependent covariates have to be categorized and forced into a multi-state model, and some information about the true relationship between the covariate and survival is bound to be lost. Another disadvantage of these multi-state models is that the number of parameters grows very quickly as more states are used. As seen in the example, dividing a continuous variable into only a few strata might not adequately describe the relationship between the covariate and survival rates, but models with more strata can introduce confounding between the different parameters. The CJS drift model solves these problems by imposing distributional constraints so that continuous variables can be included as single covariates. The number of parameters required is relatively small, and the logistic function linking the covariates and the survival rate is flexible enough to fit most monotonic relationships.

Many models of the covariate distribution could have been used, but I feel that the drift model based on the Weiner process provides a very intuitive model capturing both common trends and individual variation. I also envision many different adaptations of this model to account for a variety of different situations. Possibly the most useful adaptation would be a model in which magnitude and direction of drift is linked to the

value of the covariate itself. In its simplest form, the drift parameter on each occasion could take two values through a function like $\mu_t(z_{it}) = \mu_{1t} + \mu_{2t} I[z_{it} > z_0]$. Here z_0 might represent a cutoff between juveniles and adults so that young animals are not subject to the same forces as mature animals. Depending on the amount of data available, complex growth curves could also be incorporated so that $\mu_t(z_{it})$ takes a continuum of values. Similar models could also be used to allow for differences between male and female animals, animals living in different locations, or treatment groups in a controlled experiment. Modelling in this manner would allow researchers to use the CJS model not only to study capture and survival rates but also to understand much more about the complex relationships between different covariates.

Along with variations in the model of the covariate distribution there are many possible ways to link the covariates to the survival and capture rates. As noted by Lebreton et al. (1992), logistic regression has many nice properties for describing probabilities, but many other link functions could be used. One possibility is to use a link function based on the proportional hazards model of survival analysis. This model is designed to allow the extra flexibility of time dependent survival rates while assuming that the ratio of the instantaneous death rate for any two sets of covariates $\mathbf{z_1} \neq \mathbf{z_2}$ is a constant (Kalbfleisch and Prentice, 1980). The advantage is that time-dependence is introduced with only one extra parameter for each capture event. Exactly the same model could be used for the capture rates.

I would also like to see the application of these methods to more complex data sets. Models using two or more continuous, time-dependent, individual covariates should extend easily from this work by employing separate drift processes to describe the changes in each variable, perhaps with covariance terms linking each pair. As in Lebreton et al. (1992) indicator variables and interaction terms could be used to study the associations between the covariates and the capture and survival for different groups in the population. Discrete and continuous variables could be handled simultaneously by mixing the CJS drift and Arnason-Schwarz models. In using any of these more complex models, a researcher would need to develop suitable criteria for selecting between different subsets of the variables. The most promising measure of model fit is the Akaike Information Criterion which attempts to balance the desire for

a close fitting model as measured by the likelihood value with the need for parsimony. This is the criterion used by Lebreton et al. (1992). Other criteria might be based solely on the likelihood of different models, the deviance statistic used in the study of GLM, or Bayesian methods of model selection.

One concern that would definitely arise in using multiple covariates is that estimation seems to require very large amounts of information. This is even visible in the second simulation in which more than 300 captures occur over the 5 occasions, and yet the strong trend in survival cannot be detected. In truth, the amount of information in a data set depends on much more than just the number of captures because of the interplay between the survival and capture events. Some strange situations might arise where many captures occur but the observed covariates are grouped so that the estimated curves have very low precision except perhaps in a very small region. In higher dimensions this problem would be much worse and might prevent application of these methods except in the most ideal situations.

Another major concern is the performance of the MCEM. While the theory behind Bayesian analysis is well developed, the MCEM has not been so well explored. Theories concerning the convergence and asymptotic properties of MCEM estimates have been developed but rely on regularity conditions that may not be satisfied or more complicated sampling schemes then the one used here (Tanner, 1993). I present no guarantee that the estimates produced by averaging over several iterations of the MCEM even satisfy simple properties like consistency and unbiasedness. These issues are particularly pressing in light of the poor performance of the MCEM in analyzing the sample data set. For now, I strongly recommend use of the Bayesian estimation and feel that those unwilling to accept the Bayesian paradigm risk producing severely flawed estimates of the survival and capture rates.

Bibliography

Arnason, A. N. (1973). The estimation of population size, migration rates, and survival in a stratified population. *Research in Population Ecology* **15**, 1–8.

Borchers, D. L., Zucchini, W. and Fewster, R. M. (1998). Mark-recapture models for line transect surveys. *Biometrics* **54**, 1207–1220.

Brownie, C., Hines, J. E., Nichols, J. D., Pollock, K. H. and Hestbeck, J. B. (1993). Capture-recapture studies for multiple strata including non-markovian transitions. *Biometrics* **49**, 1173–1187.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York, 2^{nd} edition.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**, 327–335.

Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika* **51**, 429–438.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Spottiswoode, Ballantyne, and Co. Ltd., London.

Deltour, I., Richardson, S. and Le Hesran, J.-Y. (1999). Stochastic algorithms for markov models estimation with intermittent missing data. *Biometrics* **55**, 565–573.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society, Series B (Methodological)* **39**, 1–38.

Diebolt, J. and Ip, E. H. S. (1996). Stochastic EM: method and application. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 259–273. Chapman and Hall, New York.

Dupuis, J. A. (1995). Bayesian estimation of movement and survival probabilities from capture-recapture data. *Biometrika* **82**, 761–772.

Gilks, W. R. (1996). Full conditional distributions. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in Practice*, chapter 5, pages 75–88. Chapman and Hall, New York.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in Practice*, chapter 1, pages 1–19. Chapman and Hall, New York.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133–140.

Ibrahim, J. G., Chen, M.-H. and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591–596.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* **52**, 225–247.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons Ltd., Toronto.

Lebreton, J.-D., Burnham, K. P., Clobert, J. and Anderson, D. R. (1992). Modelling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecological Monographs* **62**, 67–118.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley and Sons Inc., New York.

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

Piper, S. E. (2002). Survival of adult, territorial Long-tailed Wagtail *Motacilla clara*: the effects of environmental and individual covariates. *Journal of Applied Statistics* **29**, 107–124.

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

Pollock, K. H., Hines, J. E. and Nichols, J. D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics* **40**, 329–340.

Poole, D. (2002). Bayesian estimation of survival from mark-recapture data. *Journal of Agricultural, Biological, and Environmental Statistics* **7**, 264–276.

Raftery, A. E. and Lewis, S. M. (1996). Implementing MCMC. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in Practice*, chapter 7, pages 115–130. Chapman and Hall, New York.

Schwarz, C. J., Schweigert, J. F. and Arnason, A. N. (1993). Estimating migration rates using tag-recovery data. *Biometrics* **49**, 177–193.

Seber, G. A. F. (1965). A note on the multiple recapture census. *Biometrika* **52**, 249–259.

Tanner, M. A. (1993). *Tools for Statistical Inference*. Springer-Verlag Inc., New York.

Van Deusen, P. C. (2002). An EM-algorithm for capture-recapture estimation. *Environmental and Ecological Statistics* **9**, 151–165.

Williams, B. K., Nichols, J. D. and Conroy, M. J. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego, California.